User’s Guide II

One Mean, Two Means, Cross-Over Designs,
ANOVA, Mixed Models, Multiple Comparisons,
Multivariate Means, and Microarrays

PASS

Power Analysis
and
Sample Size
System

Published by
NCSS
Dr. Jerry L. Hintze
Kaysville, Utah

PASS User's Guide II

Copyright © 2008
Dr. Jerry L. Hintze
Kaysville, Utah 84037

All Rights Reserved

Direct inquiries to:

NCSS
329 North 1000 East
Kaysville, Utah 84037
Phone (801) 546-0445
Fax (801) 546-3907
Email: support@ncss.com

PASS is a trademark of Dr. Jerry L. Hintze.

Warning:

This software and manual are both protected by U.S. Copyright Law (Title 17 United States Code). Unauthorized reproduction and/or sales may result in imprisonment of up to one year and fines of up to $10,000 (17 USC 506). Copyright infringers may also be subject to civil liability.

About This Manual

Congratulations on your purchase of the ***PASS*** package! ***PASS*** offers:

- Easy parameter entry.
- A comprehensive list of power analysis routines that are accurate and verified, yet are quick and easy to learn and use.
- Straightforward procedures for creating paper printouts and file copies of both the numerical and graphical reports.

Our goal is that with the help of these user's guides, you will be up and running on ***PASS*** quickly. After reading the quick start manual (at the front of User's Guide I) you will only need to refer to the chapters corresponding to the procedures you want to use. The discussion of each procedure includes one or more tutorials that will take you step-by-step through the tasks necessary to run the procedure.

I believe you will find that these user's guides provides a quick, easy, efficient, and effective way for first-time ***PASS*** users to get up and running.

I look forward to any suggestions you have to improve the usefulness of this manual and/or the ***PASS*** system. Meanwhile, good computing!

Jerry Hintze, Author

PASS License Agreement

Important: The enclosed Power Analysis and Sample Size software program (PASS) is licensed by NCSS to customers for their use only on the terms set forth below. Your purchase and use of the PASS system indicates your acceptance of these terms.

1. **LICENSE.** NCSS hereby agrees to grant you a non-exclusive license to use the accompanying PASS program subject to the terms and restrictions set forth in this License Agreement.

2. **COPYRIGHT.** PASS and its documentation are copyrighted. You may not copy or otherwise reproduce any part of PASS or its documentation, except that you may load PASS into a computer as an essential step in executing it on the computer and make backup copies for your use on the same computer.

3. **BACKUP POLICY.** PASS may be backed up by you for your use on the same machine for which PASS was purchased.

4. **RESTRICTIONS ON USE AND TRANSFER.** The original and any backup copies of PASS and its documentation are to be used only in connection with a single user. This user may load PASS onto several machines for his/her convenience (such as a desktop and laptop computer), but only for use by the licensee. You may physically transfer PASS from one computer to another, provided that PASS is used in connection with only one user. You may not distribute copies of PASS or its documentation to others. You may transfer this license together with the original and all backup copies of PASS and its documentation, provided that the transferee agrees to be bound by the terms of this License Agreement. PASS licenses may not be transferred more frequently than once in twelve months. Neither PASS nor its documentation may be modified or translated without written permission from NCSS.

*You may not use, copy, modify, or transfer **PASS**, or any copy, modification, or merged portion, in whole or in part, except as expressly provided for in this license.*

5. **NO WARRANTY OF PERFORMANCE.** NCSS does not and cannot warrant the performance or results that may be obtained by using PASS. Accordingly, PASS and its documentation are licensed "as is" without warranty as to their performance, merchantability, or fitness for any particular purpose. The entire risk as to the results and performance of PASS is assumed by you. Should PASS prove defective, you (and not NCSS nor its dealer) assume the entire cost of all necessary servicing, repair, or correction.

6. **LIMITED WARRANTY ON CD.** To the original licensee only, NCSS warrants the medium on which PASS is recorded to be free from defects in materials and faulty workmanship under normal use and service for a period of ninety days from the date PASS is delivered. If, during this ninety-day period, a defect in a CD should occur, the CD may be returned to NCSS at its address, or to the dealer from which PASS was purchased, and NCSS will replace the CD without charge to you, provided that you have sent a copy of your receipt for PASS. Your sole and exclusive remedy in the event of a defect is expressly limited to the replacement of the CD as provided above.

Any implied warranties of merchantability and fitness for a particular purpose are limited in duration to a period of ninety (90) days from the date of delivery. If the failure of a CD has resulted from accident, abuse, or misapplication of the CD, NCSS shall have no responsibility to replace the CD under the terms of this limited warranty. This limited warranty gives you specific legal rights, and you may also have other rights which vary from state to state.

7. **LIMITATION OF LIABILITY.** Neither NCSS nor anyone else who has been involved in the creation, production, or delivery of PASS shall be liable for any direct, incidental, or consequential damages, such as, but not limited to, loss of anticipated profits or benefits, resulting from the use of PASS or arising out of any breach of any warranty. Some states do not allow the exclusion or limitation of direct, incidental, or consequential damages, so the above limitation may not apply to you.

8. **TERM.** The license is effective until terminated. You may terminate it at any time by destroying PASS and documentation together with all copies, modifications, and merged portions in any form. It will also terminate if you fail to comply with any term or condition of this License Agreement. You agree upon such termination to destroy PASS and documentation together with all copies, modifications, and merged portions in any form.

9. **YOUR USE OF PASS ACKNOWLEDGES** that you have read this customer license agreement and agree to its terms. You further agree that the license agreement is the complete and exclusive statement of the agreement between us and supersedes any proposal or prior agreement, oral or written, and any other communications between us relating to the subject matter of this agreement.

Dr. Jerry L. Hintze & **NCSS**, Kaysville, Utah

Preface

PASS (**P**ower **A**nalysis and **S**ample **S**ize) is an advanced, easy-to-use statistical analysis software package. The system was designed and written by Dr. Jerry L. Hintze over the last Seventeen years. Dr. Hintze drew upon his experience both in teaching statistics at the university level and in various types of statistical consulting.

The present version, written for 32-bit versions of Microsoft Windows (Vista, XP, NT, ME, 2000, 98, etc.) computer systems, is the result of several iterations. Experience over the years with several different types of users has helped the program evolve into its present form.

NCSS maintains a website at www.ncss.com where we make the latest edition of ***PASS*** available for free downloading. The software is password protected, so only users with valid serial numbers may use this downloaded edition. We hope that you will download the latest edition routinely and thus avoid any bugs that have been corrected since you purchased your copy.

We believe ***PASS*** to be an accurate, exciting, easy-to-use program. If you find any portion which you feel needs to be changed, please let us know. Also, we openly welcome suggestions for additions and enhancements.

Verification

All calculations used in this program have been extensively tested and verified. First, they have been verified against the original journal article or textbook that contained the formulas. Second, they have been verified against second and third sources when these exist.

User’s Guide II
Table of Contents

Means

One Mean

Two Independent Means

Two Correlated (Paired) Means

Two Independent Means in a 2×2 Cross-Over Design

Two Independent Means in a Higher-Order Cross-Over Design

Many Means (ANOVA)

Mixed Models

Multiple Comparisons

Multivariate Means

Microarrays

References and Index

User’s Guide I
Table of Contents

Quick Start

Proportions

One Proportion

Two Correlated Proportions

Two Independent Proportions

Two Independent Proportions in a Cluster-Randomized Design

Many Proportions (Contingency Tables)

ROC Curves

References and Index

User’s Guide III
Table of Contents

Standard Deviations

Variances

One Variance

Two Variances

Normality Tests

Survival Analysis

Correlations

Regression

Design of Experiments

Tools, Helps, and Aids

References and Index

Chapter 400

Inequality Tests for One Mean (One-Sample or Paired T-Test)

Introduction

The one-sample *t* test is used to test whether the mean of a population is greater than, less than, or not equal to a specific value. Because the *t* distribution is used to calculate critical values for the test, this test is often called the one-sample *t* test. If the standard deviation is known, the normal distribution is used instead of the *t* distribution and the test is officially known as the *z test*.

When the data are differences between paired values, this test is known as the *paired t test*.

This module also calculates the power of the nonparametric analog of the *t* test, the *Wilcoxon test*.

Test Procedure

1. **Find the critical value**. Assume that the true mean is *M0*. Choose a value T_a so that the probability of rejecting H_0 when H_0 is true is equal to a specified value called α. Using the *t* distribution, select T_a so that $\Pr\left(t > T_a\right) = \alpha$. This value is found using a *t* probability table or a computer program (like ***PASS***).
2. **Select a sample of *n* items from the population and compute the *t* statistic**. Call this value *T*. If $T > T_a$ reject the null hypothesis that the mean equals *M0* in favor of an alternative hypothesis that the mean equals *M1* where *M1* > *M0*.

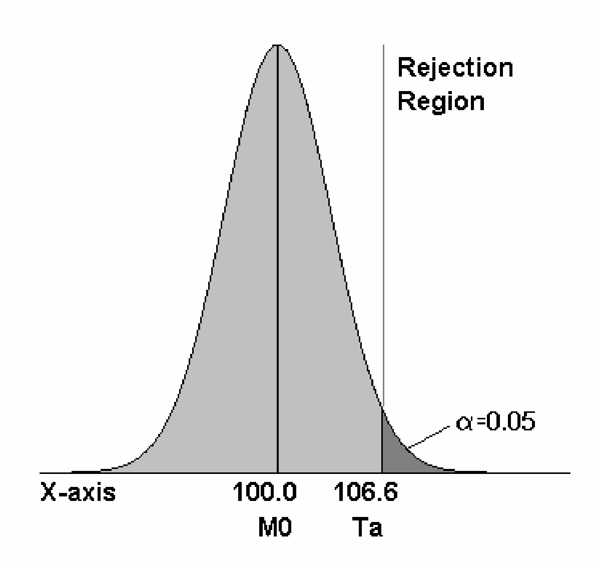
Following is a specific example. Suppose we want to test the hypothesis that a variable, *X*, has a mean of 100 versus the alternative hypothesis that the mean is greater than 100. Suppose that previous studies have shown that the standard deviation, σ, is 40. A random sample of 100 individuals is used.

We first compute the critical value, T_a. The value of T_a that yields α = 0.05 is 106.6. If the mean computed from a sample is greater than 106.6, reject the hypothesis that the mean is 100. Otherwise, do not reject the hypothesis. We call the region greater than 106.6 the *Rejection Region* and values less than or equal to 106.6 the *Acceptance Region* of the significance test.

Now suppose that you want to compute the *power* of this testing procedure. In order to compute the power, we must specify an alternative value for the mean. We decide to compute the power if the true mean were 110. Figure 2 shows how to compute the power in this case.

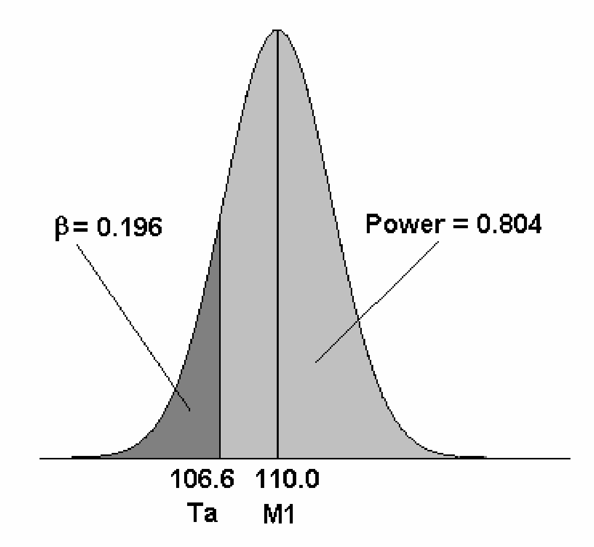
The *power* is the probability of rejecting H_0 when the true mean is 110. Since we reject H_0 when the calculated mean is greater than 106.6, the probability of a Type-II error (called β) is given by the dark, shaded area of the second graph. This value is 0.196. The power is equal to 1 - β or 0.804.

Figure 1 - Finding Alpha



Note that there are six parameters that may be varied in this situation: two means, standard deviation, alpha, beta, and the sample size.

Figure 2 - Finding Beta



Assumptions

This section describes the assumptions that are made when you use one of these tests. The key assumption relates to normality or non-normality of the data. One of the reasons for the popularity of the *t* test is its robustness in the face of assumption violation. However, if an assumption is not met even approximately, the significance levels and the power of the *t* test are invalidated. Unfortunately, in practice it often happens that several assumptions are not met. This makes matters even worse! Hence, take the steps to check the assumptions before you make important decisions based on these tests.

One-Sample T Test Assumptions

The assumptions of the one-sample *t* test are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Paired T Test Assumptions

The assumptions of the paired *t* test are:

1. The data are continuous (not discrete).
2. The data, i.e., the differences for the matched-pairs, follow a normal probability distribution.
3. The sample of pairs is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Wilcoxon Signed-Rank Test Assumptions

The assumptions of the Wilcoxon signed-rank test are as follows (note that the difference is between a data value and the hypothesized median or between the two data values of a pair):

1. The differences are continuous (not discrete).
2. The distribution of each difference is symmetric.
3. The differences are mutually independent.
4. The differences all have the same median.
5. The measurement scale is at least interval.

Limitations

There are few limitations when using these tests. Sample sizes may range from a few to several hundred. If your data are discrete with at least five unique values, you can often ignore the continuous variable assumption. Perhaps the greatest restriction is that your data come from a random sample of the population. If you do not have a random sample, your significance levels will probably be incorrect.

Technical Details

Standard Deviation Known

When the standard deviation is known, the power is calculated as follows for a directional alternative (one-tailed test) in which *M1* > *M0*.

1. Find z_α such that $1-\Phi\left(z_\alpha\right)=\alpha$, where $\Phi\left(x\right)$ is the area under the standardized normal curve to the left of *x*.

2. Calculate: $X_a = M0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.

3. Calculate: $z_a = \frac{X_a \text{ - } M1}{\frac{\sigma}{\sqrt{n}}}$.

4. Power = $1-\Phi\left(z_a\right)$.

Standard Deviation Unknown

When the standard deviation is unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which *M1* > *M0*.

1. Find t_α such that $1-T_{df}\left(t_\alpha\right)=\alpha$, where $T_{df}\left(t_\alpha\right)$ is the area under a central-*t* curve to the left of *x* and *df* = *n* - *1*.

2. Calculate: $x_a = M0 + t_\alpha \frac{\sigma}{\sqrt{n}}$.

3. Calculate the noncentrality parameter: $\lambda = \frac{M1 - M0}{\frac{\sigma}{\sqrt{n}}}$.

4. Calculate: $t_a = \frac{x_a \text{ - } M1}{\frac{\sigma}{\sqrt{n}}} + \lambda$.

5. Calculate: Power = $1-T'_{df,\lambda}\left(t_a\right)$, where $T'_{df,\lambda}\left(x\right)$ is the area under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ to the left of *x*.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power and Beta* or *N*.

Select *N* when you want to determine the sample size needed to achieve a given power and alpha error level.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study. This value must be an integer greater than one. Note that you may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Means

Mean0 (Null or Baseline)

This option specifies one or more values of the mean corresponding to the null hypothesis. If you are analyzing a paired *t* test, this value should be zero.

Only the difference between Mean0 and Mean1 is used in the calculations.

Means1 (Alternative)

This option specifies one or more values of the mean corresponding to the alternative hypothesis. If you are analyzing a paired *t* test, this value represents the mean difference that you are interested in.

Only the difference between Mean0 and Mean1 is used in the calculations.

Effect Size – Standard Deviation

Standard Deviation

This option specifies one or more values of the standard deviation. This must be a positive value. Be sure to use the standard deviation of *X* and not the standard deviation of the mean (the standard error).

When this value is not known, you must supply an estimate of it. ***PASS*** includes a special module for estimating the standard deviation. This module may be loaded by pressing the *SD* button. Refer to the Standard Deviation Estimator chapter for further details.

Known Standard Deviation

This option specifies whether the standard deviation (sigma) is known or unknown. In almost all experimental situations, the standard deviation is not known. However, great calculation efficiencies are obtained if the standard deviation is assumed to be known.

When this box is checked, the program performs its calculations assuming that the standard deviation is known. This results in the use of the normal distribution in all probability calculations. Calculations using this option will be much faster than for the unknown standard deviation case. The results for either case will be close when the sample size is over 30.

When this box is not checked, the program assumes that the standard deviation is not known and will be estimated from the data when the *t* test is run. This results in probability calculations using the noncentral-*t* distribution. This distribution requires a lot more calculations than does the normal distribution.

The calculation speed comes into play whenever the Find option is set to something besides *Beta*. In these cases, the program uses a special searching algorithm which requires numerous iterations. You will note a real difference in calculation speed depending on whether this option is checked.

A reasonable strategy would be to leave this option checked while you are experimenting with the parameters and then turn it off when you are ready for your final results.

Test

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always H_0: Mean0 = Mean1.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **Ha: Mean0 <> Mean1**

  This is the most common selection. It yields the *two-tailed t test*. Use this option when you are testing whether the means are different but you do not want to specify beforehand which mean is larger. Many scientific journals require two-tailed tests.

- **Ha: Mean0 < Mean1**

  This option yields a *one-tailed t test*. Use it when you are only interested in the case in which Mean1 is greater than Mean0.

- **Ha: Mean0 > Mean1**

  This options yields a *one-tailed t test*. Use it when you are only interested in the case in which Mean1 is less than Mean0.

Nonparametric Adjustment

This option makes appropriate sample size adjustments for the Wilcoxon test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Wilcoxon test may be made using the standard *t* test formulations with a simple adjustment to the sample size. The size of the adjustment depends upon the actual distribution of the data. They give sample size adjustment factors for four distributions. These are 1 for the uniform distribution, 2/3 for the double exponential distribution, $9/\pi^2$ for the logistic distribution, and $\pi/3$ for the normal distribution.

The options are as follows:

- **Ignore**

  Do not make a Wilcoxon adjustment. This indicates that you want to analyze a *t* test, not the Wilcoxon test.

- **Uniform**

  Make the Wilcoxon sample size adjustment assuming the uniform distribution. Since the factor is one, this option performs the same as Ignore. It is included for completeness.

- **Double Exponential**

  Make the Wilcoxon sample size adjustment assuming that the data actually follow the double exponential distribution.

- **Logistic**

  Make the Wilcoxon sample size adjustment assuming that the data actually follow the logistic distribution.

- **Normal**

  Make the Wilcoxon sample size adjustment assuming that the data actually follow the normal distribution.

Population Size

This is the number of subjects in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made.

When a finite population size is specified, the standard deviation is reduced according to the formula:

$$\sigma_1^2 = \left(1 - \frac{n}{N}\right)\sigma^2$$

where n is the sample size, N is the population size, σ is the original standard deviation, and σ_1 is the new standard deviation.

The quantity n/N is often called the sampling fraction. The quantity $\left(1 - \frac{n}{N}\right)$ is called the *finite population correction factor*.

Example 1 – Power after a Study

This example will cover the situation in which you are calculating the power of a *t* test on data that have already been collected and analyzed. For example, you might be playing the role of a reviewer, looking at the power of *t* test from a study you are reviewing. In this case, you would not vary the means, standard deviation, or sample size since they are given by the experiment. Instead, you investigate the power of the significance tests. You might look at the impact of different alpha values on the power.

Suppose an experiment involving 100 individuals yields the following summary statistics:

Hypothesized mean (M0)	100.0
Sample mean (M1)	110.0
Sample standard deviation	40.0
Sample size	100

Given the above data, analyze the power of a *t* test which tests the hypothesis that the population mean is 100 versus the alternative hypothesis that the population mean is 110. Consider the power at significance levels 0.01, 0.05, 0.10 and sample sizes 20 to 120 by 20.

Note that we have set *M1* equal to the sample mean. In this case, we are studying the power of the *t* test for a mean difference the size of that found in the experimental data.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.01 0.05 0.10**
N (Sample Size)	**20 to 120 by 20**
Mean0 (Null or Baseline)	**100**
Mean1 (Alternative)	**110**
S (Standard Deviation)	**40**
Known Standard Deviation	***Unchecked***
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
Unknown standard deviation.
                                                                     Effect
Power         N       Alpha        Beta     Mean0     Mean1        S   Size
0.06051      20     0.01000     0.93949     100.0     110.0     40.0   0.250
0.14435      40     0.01000     0.85565     100.0     110.0     40.0   0.250
0.24401      60     0.01000     0.75599     100.0     110.0     40.0   0.250
0.34953      80     0.01000     0.65047     100.0     110.0     40.0   0.250
0.45316     100     0.01000     0.54684     100.0     110.0     40.0   0.250
0.54958     120     0.01000     0.45042     100.0     110.0     40.0   0.250
0.18590      20     0.05000     0.81410     100.0     110.0     40.0   0.250
0.33831      40     0.05000     0.66169     100.0     110.0     40.0   0.250
0.47811      60     0.05000     0.52189     100.0     110.0     40.0   0.250
0.59828      80     0.05000     0.40172     100.0     110.0     40.0   0.250
0.69698     100     0.05000     0.30302     100.0     110.0     40.0   0.250
0.77532     120     0.05000     0.22468     100.0     110.0     40.0   0.250
0.28873      20     0.10000     0.71127     100.0     110.0     40.0   0.250
0.46435      40     0.10000     0.53565     100.0     110.0     40.0   0.250
0.60636      60     0.10000     0.39364     100.0     110.0     40.0   0.250
0.71639      80     0.10000     0.28361     100.0     110.0     40.0   0.250
0.79900     100     0.10000     0.20100     100.0     110.0     40.0   0.250
0.85952     120     0.10000     0.14048     100.0     110.0     40.0   0.250
```

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population. To conserve resources, it should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
Mean0 is the value of the population mean under the null hypothesis. It is arbitrary.
Mean1 is the value of the population mean under the alternative hypothesis. It is relative to Mean0.
Sigma is the standard deviation of the population. It measures the variability in the population.
Effect Size, |Mean0-Mean1|/Sigma, is the relative magnitude of the effect under the alternative.

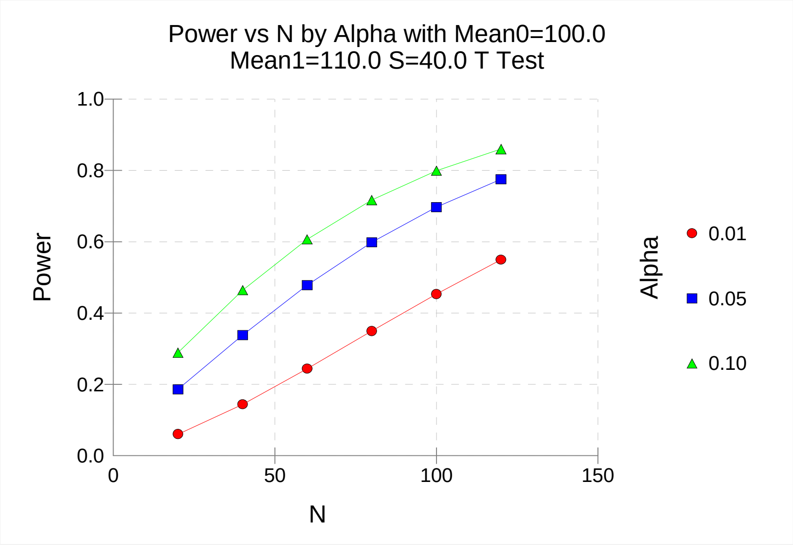
Summary Statements

A sample size of 20 achieves 6% power to detect a difference of -10.0 between the null hypothesis mean of 100.0 and the alternative hypothesis mean of 110.0 with an estimated standard deviation of 40.0 and with a significance level (alpha) of 0.01000 using a two-sided one-sample t-test.

This report shows the values of each of the parameters, one scenario per row. The values of power and beta were calculated from the other parameters.

The definitions of each column are given in the Report Definitions section.

Plots Section



This plot shows the relationship between sample size and power for various values of alpha.

Example 2 – Finding the Sample Size

This example will consider the situation in which you are planning a study that will use the one-sample *t* test and want to determine an appropriate sample size. This example is more subjective than the first because you now have to obtain estimates of all the parameters. In the first example, these estimates were provided by the data.

In studying deaths from SIDS (Sudden Infant Death Syndrome), one hypothesis put forward is that infants dying of SIDS weigh less than normal at birth. Suppose the average birth weight of infants is 3300 grams with a standard deviation of 663 grams. Use an alpha of 0.05 and power of both 0.80 and 0.90. How large a sample of SIDS infants will be needed to detect a drop in average weight of 25%? Of 10%? Of 5%? Note that applying these percentages to the average weight of 3300 yields 2475, 2970, and 3135.

Although a one-sided hypothesis is being considered, sample size estimates will assume a two-sided alternative to keep the research design in line with other studies.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80 0.90**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Mean0 (Null or Baseline)	**3300**
Mean1 (Alternative)	**2475 2970 3135**
S (Standard Deviation)	**663**
Known Standard Deviation	***Unchecked***
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.
                                                                     Effect
Power         N        Alpha         Beta     Mean0     Mean1         S     Size
0.90307       9      0.05000      0.09693    3300.0    2475.0     663.0    1.244
0.85339       8      0.05000      0.14661    3300.0    2475.0     663.0    1.244
0.90409      45      0.05000      0.09591    3300.0    2970.0     663.0    0.498
0.80426      34      0.05000      0.19574    3300.0    2970.0     663.0    0.498
0.90070     172      0.05000      0.09930    3300.0    3135.0     663.0    0.249
0.80105     129      0.05000      0.19895    3300.0    3135.0     663.0    0.249
```

This report shows the values of each of the parameters, one scenario per row. Since there were three values of Mean1 and two values of beta, there are a total of six rows in the report.

We were solving for the sample size, *N*. Notice that the increase in sample size seems to be most directly related to the difference between the two means. The difference in beta values does not seem to be as influential, especially at the smaller sample sizes.

Note that even though we set the beta values at 0.1 and 0.2, these are not the beta values that were achieved. This happens because *N* can only take on integer values. The program selects the first value of *N* that gives at least the values of alpha and beta that were desired.

Example 3 – Finding the Minimum Detectable Difference

This example will consider the situation in which you want to determine how small of a difference between the two means can be detected by the *t* test with specified values of the other parameters.

Continuing with the previous example, suppose about 50 SIDS deaths occur in a particular area per year. Using 50 as the sample size, 0.05 as alpha, and 0.20 as beta, how large of a difference between the means is detectable?

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Mean1 (Search<Mean0)**
Power	**0.80**
Alpha	**0.05**
N (Sample Size)	**50**
Mean0 (Null or Baseline)	**3300**
Mean1 (Alternative)	*Ignored since this is the Find setting*
S (Standard Deviation)	**663**
Known Standard Deviation	***Unchecked***
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.
                                                                  Effect
Power         N       Alpha        Beta    Mean0    Mean1       S    Size
0.80000      50     0.05000     0.20000   3300.0   3032.0   663.0   0.404
```

With a sample of 50, a difference of 3300 - 3032 = 268 would be detectable. This difference represents about an 8% decrease in weight.

Example 4 – Paired T Test

Usually, a researcher designs a study to compare two or more groups of subjects, so the one sample case described in this chapter occurs infrequently. However, there is a popular research design that does lead to the single mean test: *paired observations*.

For example, suppose researchers want to study the impact of an exercise program on the individual’s weight. To do so they randomly select *N* individuals, weigh them, put them through the exercise program, and weigh them again. The variable of interest is not their actual weight, but how much their weight changed.

In this design, the data are analyzed using a one-sample *t* test on the differences between the paired observations. The null hypothesis is that the average difference is zero. The alternative hypothesis is that the average difference is some nonzero value.

To study the impact of an exercise program on weight loss, the researchers decide to conduct a study that will be analyzed using the paired *t* test. A sample of individuals will be weighed before and after a specified exercise program that will last three months. The difference in their weights will be analyzed.

Past experiments of this type have had standard deviations in the range of 10 to 15 pounds. The researcher wants to detect a difference of 5 pounds or more. Alpha values of 0.01 and 0.05 will be tried. Beta is set to 0.20 so that the power is 80%. How large of a sample must the researchers take?

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

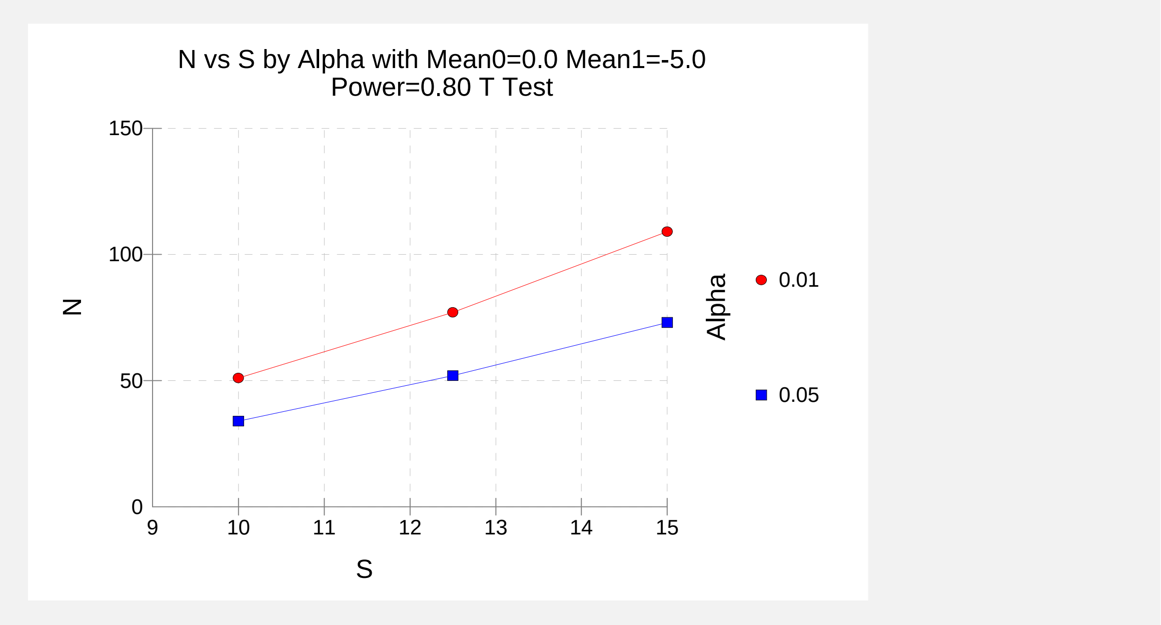
Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80**
Alpha	**0.01 0.05**
N (Sample Size)	*Ignored since this is the Find setting.*
Mean0 (Null or Baseline)	**0**
Mean1 (Alternative)	**-5**
S (Standard Deviation)	**10 12.5 15**
Known Standard Deviation	***Unchecked***
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.
                                                                  Effect
Power        N     Alpha      Beta   Mean0   Mean1        S         Size
0.80939     51   0.01000   0.19061     0.0    -5.0     10.0        0.500
0.80778     34   0.05000   0.19222     0.0    -5.0     10.0        0.500
0.80434     77   0.01000   0.19566     0.0    -5.0     12.5        0.400
0.80779     52   0.05000   0.19221     0.0    -5.0     12.5        0.400
0.80252    109   0.01000   0.19748     0.0    -5.0     15.0        0.333
0.80230     73   0.05000   0.19770     0.0    -5.0     15.0        0.333
```



The report shows the values of each of the parameters, one scenario per row. We were solving for the sample size, *N*.

Note that depending on our choice of assumptions, the sample size ranges from 34 to 109. Hence, the researchers have to make a careful determination of which standard deviation and significance level should be used.

Example 5 – Wilcoxon Test

The Wilcoxon test, a nonparametric analog of the paired comparison *t* test, is recommended when the distribution of the data is symmetrical, but not normal. A study by Al-Sunduqchi (1990) showed that sample size and power calculations for the Wilcoxon test can be made using the standard *t* test results with a simple adjustment to the sample size.

Suppose the researchers in Example 4 want to compare sample size requirements of the *t* test with those of the Wilcoxon test. They would use the same values, only this time the Nonparametric Adjustment would be set to *double exponential*. The double exponential was selected because it requires the largest adjustment of the distributions available in ***PASS*** and they wanted to know what the largest adjustment was.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

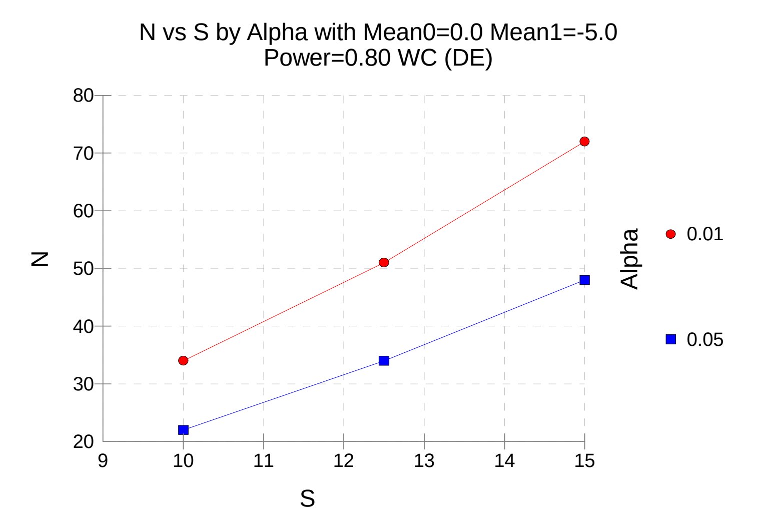
Option	Value
Data Tab	
Find (Solve For)	N
Power	0.80
Alpha	0.01 0.05
N (Sample Size)	*Ignored since this is the Find setting.*
Mean0 (Null or Baseline)	0
Mean1 (Alternative)	-5
S (Standard Deviation)	10 12.5 15
Known Standard Deviation	*Unchecked*
Alternative Hypothesis	Ha: Mean0 <> Mean1
Nonparametric Adjustment	Double Exponential
Population Size	Infinite

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.
                                                                Effect
Power        N      Alpha       Beta    Mean0    Mean1       S    Size
0.80939     34    0.01000    0.19061      0.0     -5.0    10.0   0.500
0.80778     22    0.05000    0.19222      0.0     -5.0    10.0   0.500
0.80434     51    0.01000    0.19566      0.0     -5.0    12.5   0.400
0.80779     34    0.05000    0.19221      0.0     -5.0    12.5   0.400
0.80252     72    0.01000    0.19748      0.0     -5.0    15.0   0.333
0.80230     48    0.05000    0.19770      0.0     -5.0    15.0   0.333
```



If you compare these sample size values with those of Example 4, you will find that these are about two-thirds of those required for the *t* test. This is the value of the adjustment factor for the Wilcoxon test when the underlying distribution is the double exponential.

Example 6 – Validation using Zar

Zar (1984) pages 111-112 presents an example in which Mean0 = 0.0, Mean1 = 1.0, S = 1.25, alpha = 0.05, and N = 12. Zar obtains an approximate power of 0.72.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting.*
Alpha	**0.05**
N (Sample Size)	**12**
Mean0 (Null or Baseline)	**0**
Mean1 (Alternative)	**1**
S (Standard Deviation)	**1.25**
Known Standard Deviation	***Unchecked***
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.
                                                                Effect
Power       N      Alpha       Beta    Mean0     Mean1       S    Size
0.71366    12    0.05000    0.28634      0.0       1.0     1.3   0.800
```

The difference between the power computed by ***PASS*** of 0.71366 and the 0.72 computed by Zar is mostly due to Zar's use of an approximation to the noncentral *t* distribution.

Example 7 – Validation using Machin

Machin, Campbell, Fayers, and Pinol (1997) page 37 presents an example in which Mean0 = 0.0, Mean1 = 0.2, S = 1.0, alpha = 0.05, and beta = 0.20. They obtain a sample size of 199.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (One-Sample or Paired T-Test)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (One-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example7** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Mean0 (Null or Baseline)	**0**
Mean1 (Alternative)	**0.2**
S (Standard Deviation)	**1.0**
Known Standard Deviation	***Unchecked***
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample T Test
Null Hypothesis: Mean0=Mean1     Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.
                                                             Effect
Power  N          Alpha    Beta     Mean0    Mean1    S      Size
0.80169           199      0.05000  0.19831  0.0      0.2    1.0      0.200
```

The sample size of 199 matches Machin’s result.

Chapter 405

Inequality Tests for One Exponential Mean

Introduction

This program module designs studies for testing hypotheses about the mean of the exponential distribution. Such tests are often used in *reliability acceptance testing*, also called *reliability demonstration testing*.

Results are calculated for plans that are *time censored* or *failure censored*, as well as for plans that use *with replacement* or *without replacement* sampling. We adopt the basic methodology outlined in Epstein (1960), Juran (1979), Bain and Engelhardt (1991), and Schilling (1982).

Technical Details

The test procedures described here make the assumption that lifetimes follow the exponential distribution. The density of the exponential distribution is written as

$$f(t) = \frac{1}{\theta}\exp\left(-\frac{t}{\theta}\right)$$

The parameter θ is interpreted as average failure time, mean time to failure (MTTF), or mean time between failures (MTBF). Its reciprocal is the failure rate.

The reliability, or probability that a unit continues running beyond time *t*, is

$$R(t) = e^{-\frac{t}{\theta}}$$

Hypothesis Test

The relevant statistical hypothesis is $H_0{:}\,\theta_0 = \theta_1$ versus the one-sided alternative $H_1{:}\,\theta_0 > \theta_1$. Here, θ_0 represents an acceptable (high) mean life usually set from the point of view of the producer and θ_1 represents some unacceptable (low) mean life usually set from the point of view of the consumer. The test procedure is to reject the null hypothesis if the observed mean life $\hat{\theta}$ is larger than a critical value selected to meet the error rate criterion.

The error rates are often interpreted in reliability testing as *risks*. The *consumer* runs the risk that the study will fail to reject products that have a reliability less than they have specified. This *consumer risk* is β. Similarly, the *producer* runs the risk that the study will reject products that actually meet the consumer's requirements. This *producer risk* is α.

Fixed-Failure Sampling Plans

Fixed failure plans are those in which a specified number of items, *n*, are observed until a specified number of items, r_0, fail. The length of the study t_0 is random. Failed items may, or may not, be immediately replaced (*with replacement* versus *without replacement*).

The test statistic is the observed mean life $\hat{\theta}$ which is computed using

$$\hat{\theta} = \frac{\sum\limits_{i=\text{all test items}} t_i}{r_0}$$

where t_i is the elapsed time that the *i*th item is tested, whether measured until failure or until the study is completed.

For both with-replacement and without-replacement sampling, $\hat{\theta}$ follows the two-parameter gamma distribution with density

$$g\left(y \middle| r_0, \theta\right) = \frac{1}{\left(r_0 - 1\right)!}\left(\frac{r_0}{\theta}\right)^{r} y^{r_0 - 1} e^{-r_0 y/\theta}$$

This may be converted to a standard, one-parameter gamma using the transformation

$$x = r_0 y / \theta$$

However, because chi-square tables were more accessible, and because the gamma distribution may be transformed to the chi-square distribution, most results in the statistical literature are based on the chi-square distribution. That is, $2r_0\hat{\theta}/\theta$ is distributed as a chi-square random variable with $2r_0$ degrees of freedom.

Assuming that the testing of all *n* items begins at the same instant, the expected length of time needed to observe the first r_0 failures is

$$E\left(t_0\right) = \begin{cases} \theta \sum\limits_{i=1}^{r_0} \dfrac{1}{n - i + 1} & \text{without replacement} \\ \dfrac{\theta r_0}{n} & \text{with replacement} \end{cases}$$

If you choose to solve the without replacement equation for *n*, you can make use of the approximation

$$\sum_{i=1}^{r} \frac{1}{n - i + 1} \approx \log_e\left(\frac{n + 0.5}{n - r + 0.5}\right)$$

Using the above results, sampling plans that meet the specified producer and consumer risk values may be found using the result (see Epstein (1960) page 437) that r_0 is the smallest integer such that

$$\frac{\chi^2_{\alpha,2r_0}}{\chi^2_{1-\beta,2r_0}} \geq \frac{\theta_1}{\theta_0} \text{ for testing } H_1 \colon \theta_0 > \theta_1$$

and

$$\frac{\chi^2_{\beta,2r_0}}{\chi^2_{1-\alpha,2r_0}} \geq \frac{\theta_0}{\theta_1} \text{ for testing } H_1 \colon \theta_0 < \theta_1$$

Note that the above formulation depends on r_0 but not *n*. An appropriate value of *n* can be found by considering $E(t_0)$. Two options are available.

1. The value of *n* is set (perhaps on economic grounds) and the value of $E(t_0)$ is calculated.
2. The value of $E(t_0)$ is set and the value of *n* is calculated.

Fixed-Time Sampling Plans

Fixed Time plans refer to those in which a specified number of items *n* are observed for a fixed length of time t_0. The number of items failing *r* is recorded. Sampling can be with or without replacement. The accept/reject decision can be based on *r* or the observed mean life $\hat{\theta}$ which is computed using

$$\hat{\theta} = \frac{\sum\limits_{i=\text{all test items}} t_i}{r}$$

where t_i is the time that the *i*th item is being tested, whether measured until failure or until the study is completed.

With Replacement Sampling

If failed items are immediately replaced with additional items, the distribution of *r* (and $\hat{\theta}$, since $\hat{\theta} = nt_0 / r$) follows the Poisson distribution. The probability distribution of *r* is given by the Poisson probability formula

$$P(r \leq r_0 | r, \theta) = \sum_{i=0}^{r} \frac{(nt_0 / \theta)^i}{i!} e^{-nt_0/\theta}$$

Thus, values of *n* and t_0 can be found which meet the α and β requirements.

Without Replacement Sampling

If failed items are not replaced, the distributions of *r* and $\hat{\theta}$ are different and thus the power and sample size calculations depend on which statistic will be used. The probability distribution of *r* is given by the binomial formula

$$P\left(r \le r_0 | r, \theta\right) = \sum_{i=0}^{r} \binom{n}{i} p^i \left(1-p\right)^{n-i}$$

where

$$p = 1 - e^{-t_0/\theta}$$

Thus, values of n and t_0 can be found which meet the α and β requirements. Note that this formulation ignores the actual failure times.

If $\hat{\theta}$ will be used as the test statistic, power calculations must be based on it. Bartholomew (1963) gave the following results for the case $r > 0$.

$$\Pr\left(\hat{\theta} \ge \theta_C\right) = \frac{1}{1-e^{-nt_0/\theta}} \sum_{k=1}^{n} \binom{n}{k} \sum_{i=0}^{k} \binom{k}{i} \left(-1\right)^i \exp\left\{-\frac{t_0}{\theta}\left(n-k+i\right)\right\} \int_W^{\infty} g\left(x\right) \mathrm{d}x$$

where $g(x)$ is the chi-square density function with $2k$ degrees of freedom and

$$W = \frac{2k}{\theta}\left\langle \theta_C - \frac{t_0}{k}\left(n-k+i\right)\right\rangle$$

$$\left\langle X \right\rangle = \begin{cases} X \text{ if } X > 0 \\ 0 \text{ otherwise} \end{cases}$$

The above equation is numerically unstable for large values of N, so we use the following approximation also given by Bartholomew (1961). This approximation is used when $N > 30$ or when the exact equation cannot be calculated. Bain and Engelhardt (1991) page 140 suggest that this normal approximation can be used when $p > 0.5$

$$z = \frac{u\sqrt{np}}{\sqrt{1 - \frac{2u\left(1-p\right)\log_e\left(1-p\right)}{p} + \left(1-p\right)u^2}}$$

where

$$u = \frac{\hat{\theta} - \theta}{\theta}$$

$$p = 1 - e^{-t_0/\theta}$$

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level. Select *Power and Beta* when you want to calculate the power of an experiment or test.

Error Rates

Power or Beta (Beta is Consumer's Risk)

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta (consumer's risk) is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal probabilities of the event of interest when in fact they are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Producer's Risk)

This option specifies one or more values for the probability of a type-I error (alpha), also called the producer's risk. A type-I error occurs when you reject the null hypothesis of equal probabilities when in fact they are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

Enter one or more values for the sample size N, the number of items in the study. Note that the sample size is arbitrary for sampling plans that are terminated after a fixed number of failures are observed.

You may enter a range such as 10 to 100 by 10 or a list of values separated by commas or blanks.

Test

Alternative Hypothesis

Specify the alternative hypothesis of the test. Since the null hypothesis is equality (a difference between theta0 and theta1 of zero), the alternative is all that needs to be specified. Usually, a one-tailed option is selected for these designs. In fact, the two-tailed options are only available for time terminated experiments.

Effect Size

Theta0 (Baseline Mean Life)

Enter one or more values for the *mean life* under the null hypothesis. This is sometimes called the *producer's mean life*. This value is usually scaled in terms of elapsed time such as hours, days, or years. Of course, all time values must be on the same time scale.

Note that the value of theta may be calculated from the estimated probability of failure using the relationship

$$\mathrm{P}\left(\mathrm{Failure}\right) = 1 - e^{-t0/\theta}$$

so that

$$\theta = \frac{-t0}{\ln\left(1 - P\left(\mathrm{Failure}\right)\right)}$$

Only positive values are valid. You may enter a range of values such as ‘10 20 30’ or ‘100 to 1000 by 100.’

Because the exponential function is used in the calculations, try to scale the numbers so they are less than 100. For example, instead of 720 days, use 7.2 hundreds of days. This will help to avoid numerical problems during the calculations.

Theta1 (Alternative Mean Life)

Enter one or more values for the *mean life* under the alternative hypothesis. This is sometimes called the *consumer's mean life*. This value is usually scaled in terms of elapsed time such as hours, days, or years. Of course, all time values must be on the same time scale.

Note that the value of theta may be calculated from the estimated probability of failure using the relationship

$$\mathrm{P}\left(\mathrm{Failure}\right) = 1 - e^{-t0/\theta}$$

so that

$$\theta = \frac{-t0}{\ln\left(1 - P\left(\text{Failure}\right)\right)}$$

Any positive values are valid. You may enter a range of values such as ‘10 20 30’ or ‘100 to 1000 by 100.’

Because the exponential function is used in the calculations, try to scale the numbers so they are less than 100. For example, instead of 720 days, use 7.2 hundreds of days. This will help to avoid numerical problems during the calculations.

Sampling Plan

Replacement Method

When failures occur, they may be immediately replaced (With Replacement) with new items or not (Without Replacement). One of the assumptions of the exponential distribution is that the probability of failure does not depend on the previous running time. That is, it is assumed that there is no wear-out. Adopting ‘with replacement’ sampling will shorten the elapsed time of an experiment that is failure terminated.

Termination Criterion

This option specifies the method used to terminate the study or experiment. There are two basic choices:

- **Fixed failures (*r*)**

  Terminate after *r* failures occur. This is also called *failure terminated* or *Type-II Censoring*.

- **Fixed time (*t*0)**

  Terminate after an elapsed time of *t*0. This is also called *time terminated* or *Type-I Censoring*. This is the most common.

In fixed failure sampling, *N* may be fixed while *t*0 varies or *t*0 may be fixed while *N* varies. All that matters is the product of these two quantities.

In fixed time sampling, two test statistics are available: *r* and theta-hat. When sampling is without replacement, tests based on theta-hat are more powerful (require smaller sample size).

r (Number of Failures)

Enter one or more values for the *rejection number* of the test. If *r* or more items fail, the null hypothesis that Theta0 = Theta1 is rejected in favor of the alternative the Theta0 > Theta1.

Note that this value is ignored for time terminated experiments, because the appropriate value is calculated. This value is also ignored in some situations in failure terminated experiments.

t0 (Test Duration Time)

Enter one or more values for the duration of the test. This value may be interpreted as the exact duration time, *t*0, or the expected duration time, E(*t*0), depending on the Termination Criterion and Replacement Method selected.

These values must be positive and in the same time units as Theta0 and Theta1.

E(t0) based on Theta1

When the experiment is failure terminated, the expected waiting time until *r* failures are observed, E(*t*0), is calculated. This value depends on the value of theta, the mean life. When checked, E(*t*0) calculations are based on Theta1. When unchecked, E(*t*0) calculations are based on Theta0. Either choice may be reasonable in a given situation.

Example 1 – Power for Several Sample Sizes

This example will calculate power for a time terminated, without replacement study in which the results will be analyzed using theta-hat. The study will be used to test the alternative hypothesis that Theta0 > Theta1, where Theta0 = 2.0 days and Theta1 = 1.0 days. The test duration is 1.0 days. Funding for the study will allow for a sample size of up to 40 test items. The researchers decide to look at sample sizes of 10, 20, 30, and 40. Significance levels of 0.01 and 0.05 will be considered.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Exponential Mean** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Exponential Data**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.01 0.05**
N (Sample Size)	**10 to 40 by 10**
Alternative Hypothesis	**Ha: Theta0 > Theta1**
Theta0 (Baseline Mean Life)	**2**
Theta1 (Alternative Mean Life)	**1**
Replacement Method	**Without Replacement**
Termination Criterion	**Fixed Time using Theta-hat**
t0 (Test Duration Time)	**1**

Annotated Output

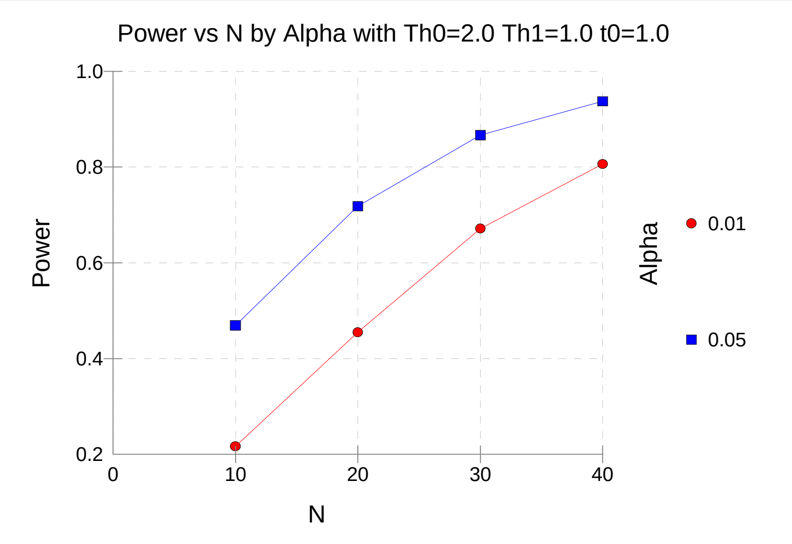
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
Test Based on Theta-hat with Fixed Running Time t0 and Without Replacement Sampling.
H0: Theta = Theta0. Ha: Theta = Theta1 < Theta0. Reject H0 if Theta-hat <= ThetaC.
                     Time                       Target     Actual   Target     Actual      Theta
Power          N       t0   Theta0   Theta1      Alpha      Alpha     Beta       Beta          C
0.21695       10    1.000      2.0      1.0    0.01000    0.01000              0.78305        0.7
0.45485       20    1.000      2.0      1.0    0.01000    0.01000              0.54515        1.0
0.67159       30    1.000      2.0      1.0    0.01000    0.01000              0.32841        1.1
0.80628       40    1.000      2.0      1.0    0.01000    0.01000              0.19372        1.2
0.46940       10    1.000      2.0      1.0    0.05000    0.05000              0.53060        1.0
0.71828       20    1.000      2.0      1.0    0.05000    0.05000              0.28172        1.2
0.86665       30    1.000      2.0      1.0    0.05000    0.05000              0.13335        1.3
0.93730       40    1.000      2.0      1.0    0.05000    0.05000              0.06270        1.4

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N is the size of the sample drawn from the population.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.
Theta0 is the Mean Life under the null hypothesis.
Theta1 is the Mean Life under the alternative hypothesis.
t0 is the test duration time. It provides the scale for Theta0 and Theta1.
r is the number of failures.

Summary Statements
A sample size of 10 achieves 22% power to detect the difference between the null hypothesis
mean lifetime of 2.0 and the alternative hypothesis mean lifetime of 1.0 at a 0.01000
significance level (alpha) using a one-sided test based on the elapsed time. Failing items are
not replaced with new items. The study is terminated when it has run for 1.000 time units.
```

This report shows the power for each of the scenarios. The critical value, Theta C, is also provided.

Plots Section



Example 2 – Validation using Epstein

Epstein (1960), page 438, presents a table giving values of *r* necessary to meet risk criteria for various values of alpha, beta, theta0, and theta1 for the fixed failures case. Specifically, when theta0 = 5, theta1 = 2, beta = 0.05, and alpha = 0.01, 0.05, and 0.10, he finds *r* = 21, 14, and 11. We will now duplicate these results.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Exponential Mean** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Exponential Data**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**r**
Power	**0.95**
Alpha	**0.01 0.05 0.10**
N (Sample Size)	**20 *(this value is ignored)***
Alternative Hypothesis	**Ha: Theta0 > Theta1**
Theta0 (Baseline Mean Life)	**5**
Theta1 (Alternative Mean Life)	**2**
Replacement Method	**Without Replacement**
Termination Criterion	**Fixed Failures, Fixed E(t0)**
t0 (Test Duration Time)	**1**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
Test Based on Fixed Failures r, Fixed Expected Time E(t0), and Without Replacement Sampling.
H0: Theta = Theta0. Ha: Theta = Theta1 < Theta0. Reject H0 if r >= r0.
                         Time                         Target     Actual    Target    Actual
Power         r0 / N    E(t0)    Theta0    Theta1      Alpha      Alpha      Beta      Beta
0.95841       21/115    1.000       5.0       2.0    0.01000    0.01000   0.05000   0.04159
0.95956        14/77    1.000       5.0       2.0    0.05000    0.05000   0.05000   0.04044
0.96221        11/60    1.000       5.0       2.0    0.10000    0.10000   0.05000   0.03779
```

PASS has calculated 21, 14, and 11 for *r* as in Epstein.

We should note that occasionally our results differ from those of Epstein. We have checked a few of these carefully by hand, and, in every case, we have found our results to be correct.

Chapter 410

Inequality Tests for One Mean (Simulation)

Introduction

This procedure allows you to study the power and sample size of several statistical tests of the hypothesis that the population mean is equal to a specific value versus the alternative that it is greater than, less than, or not equal to that value. The one-sample t-test is commonly used in this situation, but other tests have been developed for situations where the data are not normally distributed. These additional tests include the Wilcoxon signed-rank test, the sign test, and the computer-intensive bootstrap test. When the population follows the exponential distribution, a test based on this distribution should be used.

The t-test assumes that the data are normally distributed. When this assumption does not hold, the t-test is still used hoping that its robustness will produce accurate results. This procedure allows you to study the accuracy of various tests using simulation techniques. A wide variety of distributions can be simulated to allow you to assess the impact of various forms of non-normality on each test's accuracy.

The details of the power analysis of the t-test using analytic techniques are presented in the ***PASS*** chapter entitled "Inequality Tests for One Mean" and will not be duplicated here. This chapter will be confined to power analysis using computer simulation.

Technical Details

Computer simulation allows one to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. Currently, due to increased computer speeds, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are as follows.

1. Specify the method by which the test is to be carried out. This includes specifying how the test statistic is calculated and how the significance level is specified.

2. Generate a random sample, $X_1, X_2, \ldots, X_n$, from the distribution specified by the alternative hypothesis. Calculate the test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Each of these samples is used to calculate the power of the test.
3. Generate a random sample, $Y_1, Y_2, \ldots, Y_n$, from the distribution specified by the null hypothesis. Calculate the test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Each of these samples is used to calculate the significance level of the test.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data lead to a rejection of the null hypothesis. The power is the proportion of simulation samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Data Distributions

A wide variety of distributions may be studied. These distributions can vary in skewness, elongation, or other features such as bimodality. A detailed discussion of the distributions that may be used in the simulation is provided in the chapter ‘Data Simulator’.

Test Statistics

This section describes the test statistics that are available in this procedure.

One-Sample t-Test

The one-sample t-test assumes that the data are a simple random sample from a population of normally-distributed values that all have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the t-test proceeds as follows.

$$t_{n-1} = \frac{\overline{X} - M0}{s_{\overline{X}}}$$

where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

$$s_{\overline{X}} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}},$$

and $M0$ is the value of the mean hypothesized by the null hypothesis.

The significance of the test statistic is determined by computing the p-value. If this p-value is less than a specified level (often 0.05), the null hypothesis is rejected. Otherwise, no conclusion can be reached.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a popular, nonparametric substitute for the t-test. This test assumes that the data follow a symmetric distribution. The test is computed using the following steps.

1. Subtract the hypothesized mean, *M*0, from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks, *Sp*, and the sum of the negative ranks, *Sn*. The test statistic, *W*, is the minimum of *Sp* and *Sn*.
3. Compute the mean and standard deviation of *W* using the formulas

   $$\mu_W = \frac{n(n+1)}{4} \text{ and } s_W = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t_i^3 - \sum t_i}{48}}$$

   respectively, where t_i represents the number of times the i^{th} value occurs.
4. Compute the *z* value using

   $$z_W = \frac{W - \mu_W}{s_W}$$

For cases when *n* is less than 38, the significance level is found from a table of exact probabilities for the Wilcoxon test. When *n* is greater than or equal to 38, the significance of the test statistic is determined by comparing the *z* value to a normal probability table. If this p-value is less than a specified level (often 0.05), the null hypothesis is rejected. Otherwise, no conclusion can be reached.

Sign Test

The sign test is popular because it is simple to compute. This test assumes that the data all follow the same distribution. The test is computed using the following steps.

1. Count the number of values strictly greater than *M*0. Call this value *X*.
2. Count the number of values strictly less than *M*0. Call this value *Y*.
3. Set $m = X + Y$.
4. Under the null hypothesis, *X* is distributed as a binomial random variable with a proportion of 0.5 and sample size of *m*.

The significance of *X* is calculated using binomial probabilities.

Bootstrap Test

The one-sample bootstrap procedure for testing whether the mean is equal to a specific value is given in Efron & Tibshirani (1993), pages 224-227. The bootstrap procedure is as follows.

1. Compute the mean of the sample. Call it $\overline{X}$.
2. Compute the t-value using the standard t-test. The formula for this computation is

   $$t_X = \frac{\overline{X} - M0}{s_{\overline{X}}}$$

where *M0* is the hypothesized mean.

3. Draw a random, with-replacement sample of size *n* from the original *X* values. Call this sample $Y_1, Y_2, \cdots, Y_n$.
4. Compute the t-value of this bootstrap sample using the formula

$$t_Y = \frac{\overline{Y} - \overline{X}}{s_{\overline{Y}}}$$

5. For a two-tailed test, if $|t_Y| > |t_x|$ then add one to a counter variable, *A*.
6. Repeat steps 3 – 5 *B* times. *B* may be anywhere from 100 to 10,000.
7. Compute the *p*-value of the bootstrap test as (*A* + 1) / (*B* + 1)
8. Steps 1 – 7 complete one simulation iteration. Repeat these steps *M* times, where *M* is the number of simulations. The power and significance level are equal to the percent of the time the *p*-value is less than the nominal alpha of the test in their respective simulations.

Note that the bootstrap test is a time-consuming test to analyze, especially if you set *B* to a value much larger than 100.

Exponential Test

The exponential distribution is a highly skewed distribution, so it is very different from the normal distribution. Thus, the t-test does not work well with exponential data.

There is an exact test for the mean of a sample drawn from the exponential distribution. It is well known that a simple function of the mean of exponential data follows the chi-square distribution. This relationship is given in Epstein (1960) as

$$\frac{2n\overline{X}}{M0} \sim \chi^2_{2n}$$

This expression can be used to test hypotheses about the value of the mean, *M*0.

Standard Deviations

Care must be used when either the null or alternative distribution is not normal. In these cases, the standard deviation is usually not specified directly. For example, you might use a gamma distribution with a shape parameter of 1.5 and a mean of 4 as the null distribution and a gamma distribution with the same shape parameter and a mean of 5 as the alternative distribution. This allows you to compare the two means. However, note that although the shape parameters are constant, the standard deviations are not. In cases such as this, the null and alternatives not only have different means, but different standard deviations!

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated using the values of the other parameters. Under most conditions, you would select either *Power* or *N*.

Select *Power* when you want to estimate the power for a specific scenario.

Select *N* when you want to determine the sample size needed to achieve a given power and alpha error level. This option can be very computationally intensive, and may take considerable time to complete.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study. This value must be an integer greater than one. Note that you may enter a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Test

Test Type

Specify which test statistic (t-test, Wilcoxon test, sign test, bootstrap test, or exponential test) is to be simulated. Although the t-test is the most commonly used test statistic, it is based on assumptions that may not be viable in many situations. For your data, you may find that one of the other tests is more accurate (actual alpha = target alpha) and more precise (higher power).

Note that the bootstrap test is computationally intensive, so it can be very slow to evaluate.

Alternative Hypothesis

This option specifies the alternative hypothesis, H1. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always H0: Mean = M0.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **Mean <> M0**

  This is the most common selection. It yields a *two-tailed test*. Use this option when you are testing whether the mean is different from a specified value, M0, but you do not want to specify beforehand whether it is smaller or larger. Most scientific journals require two-tailed tests.

- **Mean < M0**

  This option yields a *one-tailed test*. Use it when you want to test whether the true mean is less than M0.

- **Mean > M0**

  This option yields a *one-tailed test*. Use it when you want to test whether the true mean is greater than M0.

Simulations

Simulations

This option specifies the number of iterations, M, used in the simulation. Larger numbers of iterations result in longer running time and more accurate results.

The precision of the simulated power estimates can be determined by recognizing that they follow the binomial distribution. Thus, confidence intervals may be constructed for power estimates. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95%

confidence interval of the true power is given by the power reported by the simulation plus and minus the ‘Precision’ amount given in the table.

Simulation Size M	Precision when Power = 0.50	Precision when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.014 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional precision achieved.

Effect Size

Distribution Assuming H0 (Null Hypothesis)

This option specifies the mean and distribution under the null hypothesis, H0. Usually, the mean is specified by entering ‘M0’ for the mean parameter in the distribution expression and then entering values for the M0 parameter described below. All of the distributions are parameterized so that the mean is entered first. For example, if you wanted to test whether the mean of a normal distributed variable is five, you could enter N(5, S) or N(M0, S) here.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value ‘M0’ is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them:

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,P3,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean under H0

The distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Specifying the Mean for Paired (Matched) Data

Depending on the formula that is entered, the mean is not necessarily the value of M0. For example, a common use of the one-group t-test is to test whether the mean of a set of differences is zero. Differences may be specified (ignoring the correlation between paired observations) as the difference between two normal distributions. This would be specified as N(M0, S) - N(M0, S). The mean of the resulting distribution is M0 – M0 = 0 (not M0).

Distribution Assuming H1 (Alternative Hypothesis)

This option specifies the mean and distribution under the alternative hypothesis, H1. That is, this is the actual (true) value of the mean at which the power is computed. Usually, the mean is specified by entering ‘M1’ for the mean parameter in the distribution expression and then entering values for the M1 parameter below. All of the distributions are parameterized so that the mean is entered first.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value ‘M1’ is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them:

```
Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,P3,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)
```

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean under H1

The distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Specifying the Mean for Paired (Matched) Data

Depending on the formula that is entered, the mean is not necessarily the value of M1. For example, a common use of the one-group t-test is to test whether the mean of a set of differences is zero. Differences may be specified (ignoring the correlation between paired observations) as the difference between two normal distributions. This would be specified as N(M1, S) - N(M0, S). The mean of the resulting distribution is M1 – M0.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for the M0 in the distribution specifications given above. M0 is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using syntax such as *0 1 2 3* or *0 to 3 by 1*.

Note that whether M0 is the mean of the simulated distribution depends on the formula you have entered. For example, N(M0, S) has a mean of M0, but N(M0, S)-N(M0, S) has a mean of zero.

M1 (Mean|H1)

These values are substituted for the M1 in the distribution specifications given above. M1 is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using syntax such as *0 1 2 3* or *0 to 3 by 1*.

Note that whether M1 is the mean of the simulated distribution depends on the formula you have entered. For example, N(M1, S) has a mean of M1, but N(M1, S)-N(M0, S) has a mean of M1 - M0.

Parameter Values (S, A, B, C)

Enter the numeric value(s) of parameter listed above. These values are substituted for the corresponding letter in the distribution specifications for H0 and H1.

You can enter a list of values using syntax such as *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter.

Iterations Tab

The Iterations tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size, N, is aborted. When the maximum number of iterations is reached without convergence, the sample size is not reported. We recommend a value of at least 500.

Bootstrap Iterations

Bootstrap Iterations

Specify the number of iterations used in the bootstrap hypothesis test. This value is only used if the bootstrap test is displayed on the reports. The running time of the procedure depends heavily on the number of iterations specified here.

Recommendations by authors of books discussing the bootstrap range from 100 to 10,000. If you enter a large (greater than 500) value, the procedure may take several hours to run.

Example 1 – Power at Various Sample Sizes

A researcher is planning an experiment to test whether the mean response level to a certain drug is significantly different from zero. The researcher wants to use a t-test with an alpha level of 0.05. He wants to compute the power at various sample sizes from 5 to 40, assuming the true mean is one. He assumes that the data are normally distributed with a standard deviation of 2. Since this is an exploratory analysis, he sets the number of simulation iterations to 1000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**5 to 40 by 5**
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**1000**
Distribution|H0 (Null Hypothesis)	**N(M0 S)**
Distribution|H1 (Alt Hypothesis)	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**2**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc’s & 95% C.I.’s	**Checked**
Show Definitions	**Checked**
Show Plots	**Checked**
Number of Summary Statements	**1**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

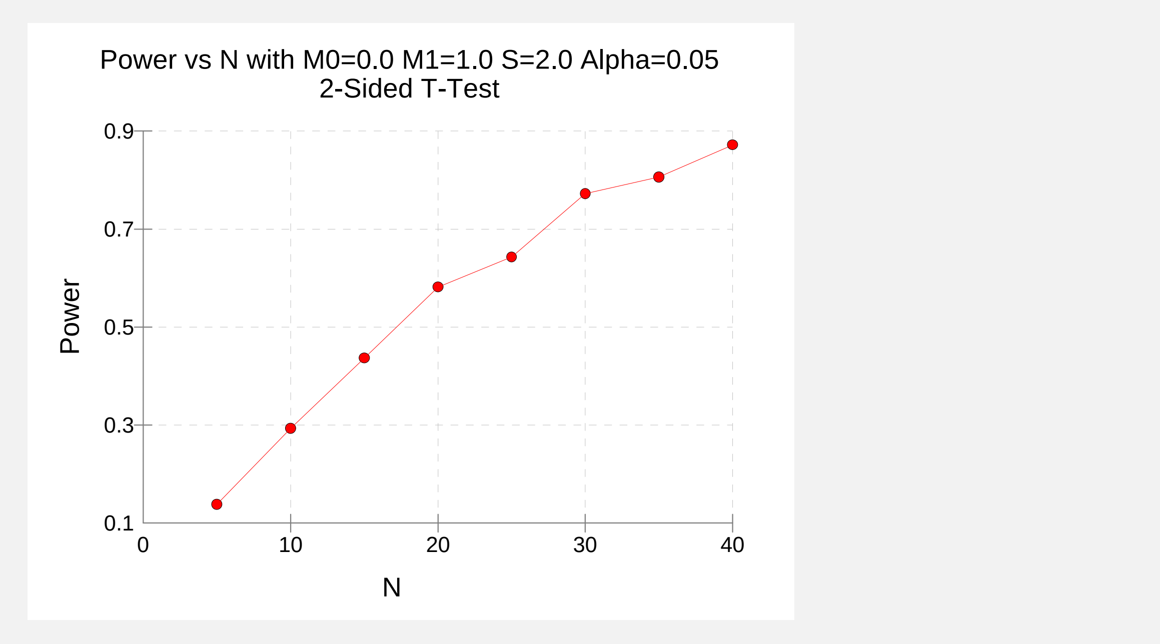
Numeric Results

```
Numeric Results for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Normal(M0 S)
H1 Distribution: Normal(M1 S)
Test Statistic: T-Test

                      H0       H1   Target    Actual
Power         N    Mean0    Mean1    Alpha     Alpha      Beta        M0        M1         S
0.138         5      0.0      1.0    0.050     0.050     0.862       0.0       1.0       2.0
(0.021)  [0.117   0.159]                     (0.014)    [0.036    0.064]

0.293        10      0.0      1.0    0.050     0.061     0.707       0.0       1.0       2.0
(0.028)  [0.265   0.321]                     (0.015)    [0.046    0.076]

0.437        15      0.0      1.0    0.050     0.058     0.563       0.0       1.0       2.0
(0.031)  [0.406   0.468]                     (0.014)    [0.044    0.072]

0.582        20      0.0      1.0    0.050     0.058     0.418       0.0       1.0       2.0
(0.031)  [0.551   0.613]                     (0.014)    [0.044    0.072]

0.643        25      0.0      1.0    0.050     0.048     0.357       0.0       1.0       2.0
(0.030)  [0.613   0.673]                     (0.013)    [0.035    0.061]

0.772        30      0.0      1.0    0.050     0.042     0.228       0.0       1.0       2.0
(0.026)  [0.746   0.798]                     (0.012)    [0.030    0.054]

0.806        35      0.0      1.0    0.050     0.054     0.194       0.0       1.0       2.0
(0.025)  [0.781   0.831]                     (0.014)    [0.040    0.068]

0.872        40      0.0      1.0    0.050     0.044     0.128       0.0       1.0       2.0
(0.021)  [0.851   0.893]                     (0.013)    [0.031    0.057]

Notes:
Second Row: (Power Inc.) [95% LCL and UCL Power]    (Alpha Inc.) [95% LCL and UCL Alpha]
Number of Monte Carlo Samples: 1000.   Simulation Run Time: 17.81 seconds.

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N is the size of the sample drawn from the population.
Mean0 is the value of the mean assuming the null hypothesis. This is the value being tested.
Mean1 is the actual value of the mean. The procedure tests whether Mean0 = Mean1.
Target Alpha is the probability of rejecting a true null hypothesis. It is set by the user.
Actual Alpha is the alpha level that was actually achieved by the experiment.
Beta is the probability of accepting a false null hypothesis.
```

This report shows the estimated power for each scenario. The first row shows the parameter settings and the estimated power and significance level (Actual Alpha). Note that because these are results of a simulation study, the computed power and alpha will vary from run to run. Thus, another report obtained using the same input parameters will be slightly different than the one above.

The second row shows two 95% confidence intervals in brackets: the first for the power and the second for the significance level. Half the width of each confidence interval is given in parentheses as a fundamental measure of the accuracy of the simulation. As the number of simulations is increased, the width of the confidence interval will decrease.

Plots Section



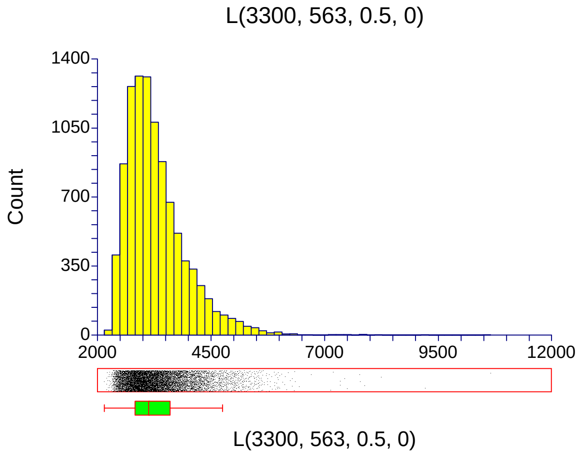
This plot shows the relationship between sample size and power.

Example 2 – Finding the Sample Size for Skewed Data

In studying deaths from SIDS (Sudden Infant Death Syndrome), one hypothesis put forward is that infants dying of SIDS weigh less than normal at birth. Suppose the average birth weight of infants is 3300 grams with a standard deviation of 663 grams. The researchers decide to examine the effect of a skewed distribution on the test used by adding skewness to the simulated data using Tukey's Lambda distribution with a skewness factor of 0.5.

Using the Data Simulator program, the researchers found that the actual standard deviation using the above parameters was almost 800. This occurs because adding skewness changes the standard deviation. They found that setting the standard deviation in Tukey's Lambda distribution to 563 resulted in a standard deviation in the data of about 663.

A histogram of 10,000 pseudo-random values from this distribution appears as follows.



The researchers want to determine how large a sample of SIDS infants will be needed to detect a drop in average weight of 25%? Note that applying this percentage to the average weight of 3300 yields 2475.Use an alpha of 0.05 and 80% power.

Although a one-sided hypothesis might be considered, sample size estimates will assume a two-sided alternative to keep the research design in line with other studies. To decrease the running time of this example, the number of simulation iterations is set to 1000. In practice, you would probably use a value of about 5000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**1000**
Distribution|H0 (Null Hypothesis)	**L(M0 S G 0)**
Distribution|H1 (Alt Hypothesis)	**L(M1 S G 0)**
M0 (Mean|H0)	**2475**
M1 (Mean|H1)	**3300**
S	**563**
G	**0.5** (Note that parameter A was changed to G.)
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results of Search for N

```
                         H0        H1    Target    Actual
Power          N      Mean0     Mean1     Alpha     Alpha       Beta        M0        M1          S
0.817          6     2475.0    3300.0     0.050     0.073      0.183    2475.0    3300.0      563.0        0.5
(0.024)   [0.793     0.841]                        (0.016)    [0.057     0.089]
```

The required sample size was 6. Notice how wide the confidence interval of power is. We re-ran this simulation several times and obtained sample sizes of 5, 6, and 7. Note that the actual alpha value is between 0.057 and 0.089, which is definitely greater than 0.05. This shows one of the problems of using the t-test with a skewed distribution.

To be more accurate and yet avoid the long running time of the search for N, a reasonable strategy would be to run simulations to obtain the powers using N’s from 4 to 10. The result of this study is displayed next.

Numeric Results of Power Search for Various N

```
                    H0       H1  Target  Actual
Power        N   Mean0    Mean1   Alpha   Alpha     Beta      M0      M1        S
0.414        4  2475.0   3300.0   0.050   0.093    0.586  2475.0  3300.0    563.0       0.5
(0.014) [0.400   0.428]                  (0.008)  [0.085   0.101]

0.645        5  2475.0   3300.0   0.050   0.084    0.355  2475.0  3300.0    563.0       0.5
(0.013) [0.632   0.658]                  (0.008)  [0.076   0.091]

0.811        6  2475.0   3300.0   0.050   0.088    0.189  2475.0  3300.0    563.0       0.5
(0.011) [0.800   0.822]                  (0.008)  [0.081   0.096]

0.912        7  2475.0   3300.0   0.050   0.089    0.088  2475.0  3300.0    563.0       0.5
(0.008) [0.905   0.920]                  (0.008)  [0.081   0.097]

0.960        8  2475.0   3300.0   0.050   0.077    0.040  2475.0  3300.0    563.0       0.5
(0.005) [0.955   0.966]                  (0.007)  [0.069   0.084]

0.983        9  2475.0   3300.0   0.050   0.082    0.017  2475.0  3300.0    563.0       0.5
(0.004) [0.979   0.987]                  (0.008)  [0.074   0.089]

0.994       10  2475.0   3300.0   0.050   0.079    0.006  2475.0  3300.0    563.0       0.5
(0.002) [0.992   0.996]                  (0.007)  [0.071   0.086]
```

The sample size of 6 appears to meet the design parameters the best. The actual significance level still appears to be greater than 0.05. The researchers decide that they must use a smaller value of Alpha so that the actual alpha is about 0.05. After some experimentation, they find that setting Alpha to 0.025 results in the desired power and significance level.

Numeric Results with Alpha = 0.025

```
                    H0       H1  Target  Actual
Power        N   Mean0    Mean1   Alpha   Alpha     Beta      M0      M1        S
0.593        6  2475.0   3300.0   0.025   0.057    0.407  2475.0  3300.0    563.0       0.5
(0.014) [0.579   0.606]                  (0.006)  [0.051   0.064]

0.754        7  2475.0   3300.0   0.025   0.058    0.246  2475.0  3300.0    563.0       0.5
(0.012) [0.742   0.766]                  (0.006)  [0.051   0.064]

0.862        8  2475.0   3300.0   0.025   0.049    0.138  2475.0  3300.0    563.0       0.5
(0.010) [0.853   0.872]                  (0.006)  [0.043   0.055]

0.929        9  2475.0   3300.0   0.025   0.044    0.071  2475.0  3300.0    563.0       0.5
(0.007) [0.921   0.936]                  (0.006)  [0.039   0.050]
```

It appears that a sample size of 8 with a Target Alpha of 0.025 will result in an experimental design with the characteristics the researchers wanted.

Notice that when working with non-normal distributions, you must change both N and the Target Alpha to achieve the design you want!

Example 3 – Comparative results with Skewed Data

Continuing with Example2, the researchers want to study the characteristics of various test statistics as the amount of skewness is increased. To do this, they let the skewness parameter of Tukey's Lambda distribution vary between 0 and 1. The researchers realize that the standard deviation will change as the skewness parameter is increased, but they decide to ignore this complication.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**6**
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**1000**
Distribution|H0 (Null Hypothesis)	**L(M0 S G 0)**
Distribution|H1 (Alt Hypothesis)	**L(M1 S G 0)**
M0 (Mean|H0)	**2475**
M1 (Mean|H1)	**3300**
S	**563**
G	**0.0 0.2 0.4 0.6 0.8 1.0**
Report Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**
Include T-Test Results	**Checked**
Include Wilcoxon & Sign Test	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing One Mean = Mean0.   Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Tukey(M0 S G 0)
H1 Distribution: Tukey(M1 S G 0)
              H0        H1
            Mean      Mean    Target   T-TestWilcoxon      Sign
N        (Mean0)   (Mean1)     Alpha    Power    Power     Power
6         2475.0    3300.0     0.050    0.816    0.634     0.634
6         2475.0    3300.0     0.050    0.852    0.705     0.705
6         2475.0    3300.0     0.050    0.845    0.790     0.790
6         2475.0    3300.0     0.050    0.779    0.839     0.839
6         2475.0    3300.0     0.050    0.644    0.866     0.866
6         2475.0    3300.0     0.050    0.466    0.757     0.757
Number of Monte Carlo Iterations: 5000.   Simulation Run Time: 43.81 seconds.

Alpha Comparison for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
              H0        H1
            Mean      Mean    Target   T-TestWilcoxon      Sign
N        (Mean0)   (Mean1)     Alpha    Alpha    Alpha     Alpha
6         2475.0    3300.0     0.050    0.046    0.032     0.032
6         2475.0    3300.0     0.050    0.058    0.035     0.035
6         2475.0    3300.0     0.050    0.070    0.040     0.040
6         2475.0    3300.0     0.050    0.095    0.056     0.056
6         2475.0    3300.0     0.050    0.134    0.084     0.084
6         2475.0    3300.0     0.050    0.173    0.107     0.107
Number of Monte Carlo Iterations: 5000.   Simulation Run Time: 43.81 seconds.
```

Several interesting trends become apparent from this study. First, for a sample size of 6, the power of the Wilcoxon test and the sign test are the same (this is not the case for larger sample sizes). The power of the t-test decreases as the amount of skewness increases. Unfortunately, we do not know if this was due to the increased variance, or the increased skewness. The power of the Wilcoxon and sign tests does not decrease—in fact, it increases until the skewness reaches 0.6. Finally, the significance level is adversely impacted by the skewness.

Example 4 – Validation using Zar

Zar (1984), pages 111-112, presents an example in which Mean0 = 0.0, Mean1 = 1.0, S = 1.25, alpha = 0.05, and N = 12. Zar obtains an approximate power of 0.72. We will validate this procedure by running this example. To make certain that the results are very accurate, the number of simulations will be set to 10,000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**12**
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**10000**
Distribution|H0 (Null Hypothesis)	**N(M0 S)**
Distribution|H1 (Alt Hypothesis)	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**1.25**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc’s & 95% C.I.’s	**Checked**
Show Definitions	**Checked**
Show Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Normal(M0 S)
H1 Distribution: Normal(M1 S)
Test Statistic: T-Test
                          H0        H1    Target    Actual
Power          N       Mean0     Mean1     Alpha     Alpha      Beta        M0        M1         S
0.717         12         0.0       1.0     0.050     0.056     0.283       0.0       1.0       1.3
(0.009)   [0.708      0.726]                       (0.004)    [0.051     0.060]
```

This simulation obtained a power of 0.717 which rounds to the 0.72 computed by Zar. Note that another repetition of this same analysis will probably be slightly different since a different set of random numbers will be used.

Example 5 – Validation using Machin

Machin, et. al. (1997), page 37, present an example in which Mean0 = 0.0, Mean1 = 0.2, S = 1.0, alpha = 0.05, and beta = 0.20. They obtain a sample size of 199. Because of the long running time, we will set the number of simulations at only 200. Of course, in practice you would usually set this to a value greater than 1000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**200**
Distribution|H0 (Null Hypothesis)	**N(M0 S)**
Distribution|H1 (Alt Hypothesis)	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.20**
S	**1**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Normal(M0 S)
H1 Distribution: Normal(M1 S)
Test Statistic: T-Test
                        H0         H1    Target    Actual
Power         N      Mean0      Mean1     Alpha     Alpha      Beta        M0        M1         S
0.785       211        0.0        0.2     0.050     0.045     0.215       0.0       0.2       1.0
(0.057)  [0.728      0.842]                       (0.029)    [0.016    0.074]

Notes:
Second Row: (Power Inc.) [95% LCL and UCL Power]    (Alpha Inc.) [95% LCL and UCL Alpha]
Number of Monte Carlo Samples: 200.   Simulation Run Time: 39.83 seconds.
```

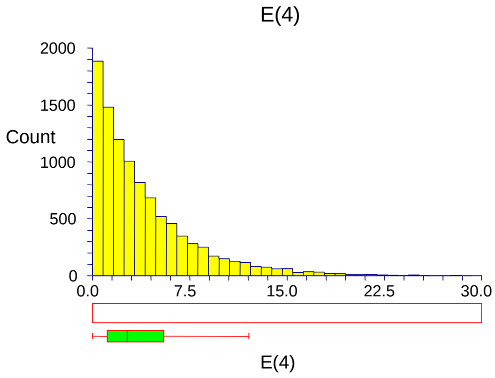
Note that using a simulation size of only 200, the estimated sample size of 211 is still close to the exact value of 199. We ran this simulation several times and obtained sample sizes between 187 and 211.

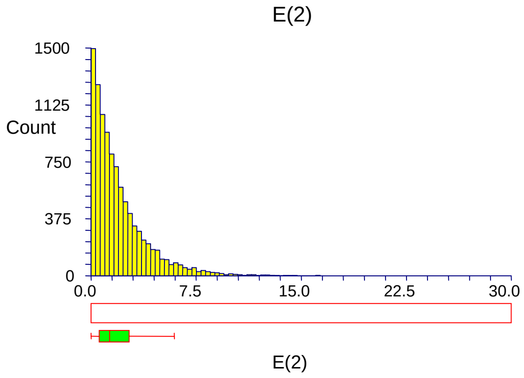
You might try resetting the simulation size to 2000 and rerunning the simulation.

Example 6 – Power of the Wilcoxon Test

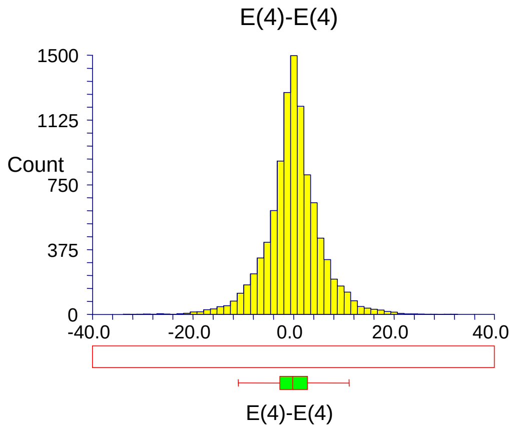
The Wilcoxon nonparametric test was designed for data that do not follow the normal distribution but are symmetric. This type of data often occurs when differences between two non-normal variables are taken, as in a study that analyzes differences in pre- and post-test scores.

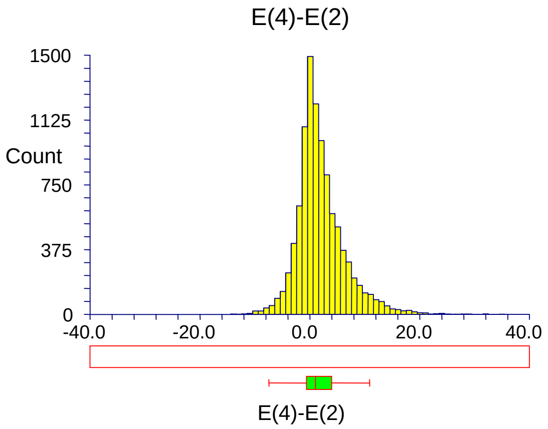
For this example, suppose the pre-test and the post-test scores are exponentially distributed. Here are examples of exponentially-distributed data with means of 4 and 2, respectively.





It has been shown that the differences between two identically-distributed variables are symmetric. The histogram below on the left shows differences in the null case in which the difference is between two exponential variables both with a mean of 4. The histogram below on the right shows differences in the alternative case in which the difference is between an exponential variable with a mean of 4 and an exponential variable with a mean of 2. Careful inspection shows that the second histogram is skewed to the right and the mean difference is about 2, not 0.





The researchers want to study the power of the two-sided Wilcoxon test when sample sizes of 10, 20, 30, and 40 are used, and testing is done at the 5% significance level.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**10 20 30 40 50**
Test Type	**Wilcoxon**
Alternative Hypothesis	**Mean<>M0**
Simulations	**5000**
Distribution|H0 (Null Hypothesis)	**E(M0)-E(M0)**
Distribution|H1 (Alt Hypothesis)	**E(M0)-E(M1)**
M0 (Mean|H0)	**4**
M1 (Mean|H1)	**2**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc’s & 95% C.I.’s	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Expo(M0)-Expo(M0)
H1 Distribution: Expo(M0)-Expo(M1)
Test Statistic: Wilcoxon Signed-Rank Test

                     H0       H1   Target    Actual
Power        N    Mean0    Mean1    Alpha     Alpha      Beta        M0         M1
0.204       10      0.0      2.0    0.050     0.038     0.796       4.0        2.0
(0.011)  [0.193   0.216]                    (0.005)    [0.032    0.043]

0.480       20      0.0      2.0    0.050     0.051     0.520       4.0        2.0
(0.014)  [0.466   0.494]                    (0.006)    [0.045    0.057]

0.647       30      0.0      2.0    0.050     0.050     0.353       4.0        2.0
(0.013)  [0.634   0.660]                    (0.006)    [0.044    0.056]

0.789       40      0.0      2.0    0.050     0.047     0.211       4.0        2.0
(0.011)  [0.778   0.800]                    (0.006)    [0.041    0.053]

0.863       50      0.0      2.0    0.050     0.049     0.137       4.0        2.0
(0.010)  [0.853   0.872]                    (0.006)    [0.043    0.055]

Notes:
Second Row: (Power Inc.) [95% LCL and UCL Power]    (Alpha Inc.) [95% LCL and UCL Alpha]
Number of Monte Carlo Samples: 5000.   Simulation Run Time: 79.70 seconds.
```

Reasonable power is achieved for N = 50.

Example 7 – Likert-Scale Data

Likert-scale data occurs commonly in survey research. A *Likert Scale* is discrete, ordinal data. It usually occurs when a survey poses a question and the respondent must pick among strongly agree, agree, undecided, disagree, or strongly disagree. The responses are usually coded as 1, 2, 3, 4, and 5.

Likert data can be analyzed in a number of ways. Perhaps the most common is to use a t-test or a Wilcoxon test. (Using the Wilcoxon test is invalid in this case because the data are seldom distributed symmetrically.)

In this example, a questionnaire is planned on which Likert-scale questions will be asked. The researchers want to study the power and actual significance levels of various sample sizes. They decide to look at what happens as the proportion of strongly agree responses is increased beyond a perfectly uniform response pattern. They want to compute the power when the strongly agree response is twice as likely, four times as likely, and eight times as likely. The sample size is 20, alpha is 0.05, and the test is two-sided.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using**

Differences (Simulation). You may then follow along here by making the appropriate entries as listed below or load the completed template **Example7** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**20**
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**5000**
Distribution|H0 (Null Hypothesis)	**M(M0 1 1 1 1)**
Distribution|H1 (Alt Hypothesis)	**M(M1 1 1 1 1)**
M0 (Mean|H0)	**1**
M1 (Mean|H1)	**2 4 8**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: M(M0 1 1 1 1)
H1 Distribution: M(M1 1 1 1 1)
Test Statistic: T-Test

                          H0       H1   Target   Actual
Power          N       Mean0    Mean1    Alpha    Alpha      Beta        M0        M1
0.167         20         3.0      2.7    0.050    0.050     0.833       1.0       2.0
(0.010)  [0.156        0.177]                   (0.006)    [0.044    0.056]

0.558         20         3.0      2.3    0.050    0.052     0.442       1.0       4.0
(0.014)  [0.544        0.572]                   (0.006)    [0.046    0.058]

0.910         20         3.0      1.8    0.050    0.055     0.090       1.0       8.0
(0.008)  [0.902        0.918]                   (0.006)    [0.048    0.061]

Notes:
Second Row: (Power Inc.) [95% LCL and UCL Power]    (Alpha Inc.) [95% LCL and UCL Alpha]
Number of Monte Carlo Samples: 5000.   Simulation Run Time: 12.53 seconds.
```

Note that M0 and M1 are no longer the H0 and H1 means. Now, they represent the relative weighting given to the strongly agree response. Under H0, the mean is 3.0. As M1 is increased, the mean under H1 changes from 2.7 to 2.3 to 1.8. We note that the actual significance level, alpha, remains close to the target value of 0.05.

Example 8 – Computing the Power after Completing an Experiment

A group of researchers has completed an experiment designed to determine if a particular hormone increases weight gain in rats. The researchers inject 20 rats of the same age with the hormone and measure their weight gain after 1 month. The investigators uses the two-sided bootstrap test with alpha = 0.05 and 100 bootstrap samples to determine if the average weight gained by these rats (171 grams) is significantly greater than the known average weight gained by rats of the same age over the same period of time (155 grams). Unfortunately, the results indicate that there is no significant difference between the two means. Therefore, the researchers decide to compute the power achieved by this test for alternative means ranging from 160 to 190 grams. They decide to use 1000 simulations for the study. For comparative purposes, they also decide to look at the power achieved by the bootstrap test in comparison to various other applicable tests. Suppose that they know that the standard deviation for weight gain is 33 grams.

Note that the researchers compute the power for a range of practically significant alternatives. The range chosen should represent likely values based on historical evidence.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for One Mean (Simulation)** procedure window by clicking on **Means**, then **One Mean**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example8** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**20**
Test Type	**Bootstrap**
Alternative Hypothesis	**Mean<>M0**
Simulations	**1000**
Distribution|H0 (Null Hypothesis)	**N(M0 S)**
Distribution|H1 (Alt Hypothesis)	**N(M1 S)**
M0 (Mean|H0)	**155**
M1 (Mean|H1)	**160 to 190 by 10**
S	**33**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**
Show Comparative Reports	**Checked**
Show Plots	**Checked**
Show Comparative Plots	**Checked**
Include T-Test Results	**Checked**

Reports Tab (continued)

Include Wilcoxon & Sign Test .... **Checked**

Include Bootstrap Test Results .... **Checked**

Iterations Tab

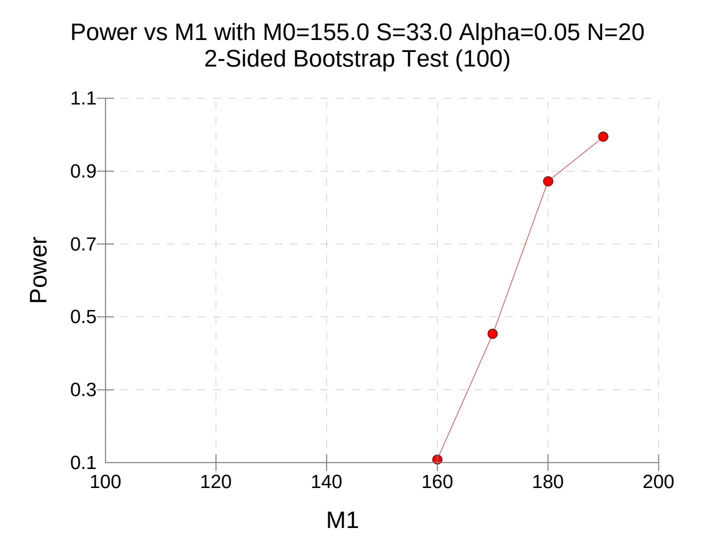
Bootstrap Iterations **100**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results for Power of Bootstrap

```
Numeric Results for Testing One Mean = Mean0.    Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Normal(M0 S)
H1 Distribution: Normal(M1 S)
Test Statistic: Bootstrap Test (100)

                      H0        H1   Target    Actual
Power        N     Mean0     Mean1    Alpha     Alpha      Beta        M0        M1         S
0.108       20     155.0     160.0    0.050     0.045     0.892     155.0     160.0      33.0
(0.019)  [0.089    0.127]                     (0.013)    [0.032    0.058]

0.453       20     155.0     170.0    0.050     0.044     0.547     155.0     170.0      33.0
(0.031)  [0.422    0.484]                     (0.013)    [0.031    0.057]

0.872       20     155.0     180.0    0.050     0.044     0.128     155.0     180.0      33.0
(0.021)  [0.851    0.893]                     (0.013)    [0.031    0.057]

0.994       20     155.0     190.0    0.050     0.042     0.006     155.0     190.0      33.0
(0.005)  [0.989    0.999]                     (0.012)    [0.030    0.054]

Notes:
Second Row: (Power Inc.) [95% LCL and UCL Power]    (Alpha Inc.) [95% LCL and UCL Alpha]
Number of Monte Carlo Samples: 1000.   Simulation Run Time: 2.99 minutes.
```



Reasonable power is achieved by this test for alternative means larger than 180. The accuracy of these results, of course, depends on the assumption that the data are normally distributed.

Comparative Results for Power of Various Tests

Power Comparison for Testing One Mean = Mean0.   Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Normal(M0 S)
H1 Distribution: Normal(M1 S)

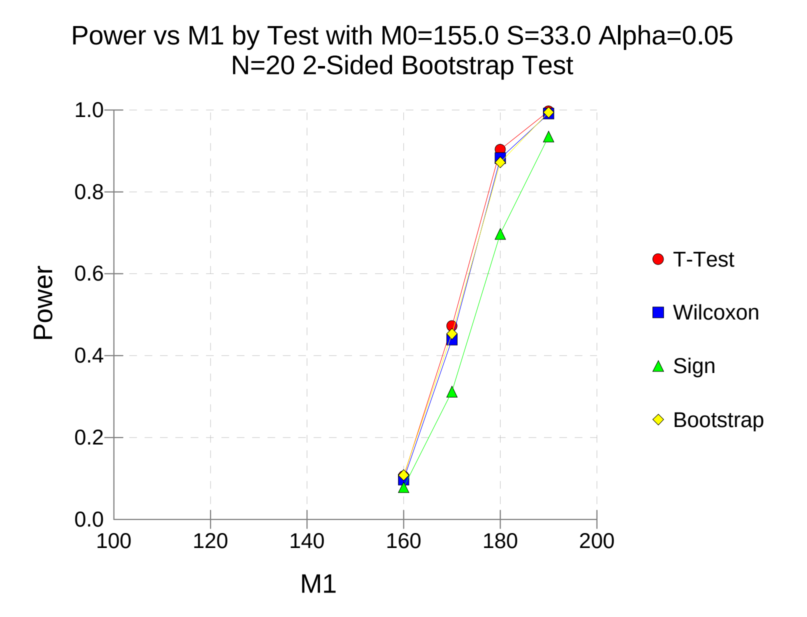
	H0	H1					
	Mean	Mean	Target	T-Test	Wilcoxon	Sign	Bootstrap
N	(Mean0)	(Mean1)	Alpha	Power	Power	Power	Power
20	155.0	160.0	0.050	0.105	0.097	0.078	0.108
20	155.0	170.0	0.050	0.472	0.439	0.312	0.453
20	155.0	180.0	0.050	0.903	0.882	0.697	0.872
20	155.0	190.0	0.050	0.997	0.991	0.935	0.994

Number of Monte Carlo Iterations: 1000.   Simulation Run Time: 2.99 minutes.

Alpha Comparison for Testing One Mean = Mean0.   Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Normal(M0 S)
H1 Distribution: Normal(M1 S)

	H0	H1					
	Mean	Mean	Target	T-Test	Wilcoxon	Sign	Bootstrap
N	(Mean0)	(Mean1)	Alpha	Alpha	Alpha	Alpha	Alpha
20	155.0	160.0	0.050	0.041	0.045	0.038	0.045
20	155.0	170.0	0.050	0.045	0.049	0.037	0.044
20	155.0	180.0	0.050	0.045	0.040	0.033	0.044
20	155.0	190.0	0.050	0.053	0.058	0.037	0.042

Number of Monte Carlo Iterations: 1000.   Simulation Run Time: 2.99 minutes.



It is apparent from these results that the bootstrap performs as well as (if not better than) the t-test and nonparametric tests for this design.

Example 9 – Comparison of Tests for Exponential Data

A researcher is designing an experiment. She believes that the data will follow an exponential distribution. Consequently, she does not believe that the t-test will be useful for her situation. She would like to compare several possible tests to determine which would be best for analyzing exponential data. She is interested in determining the power when the alternative mean is twice the null mean, which is 10. She wants to find the power achieved for sample sizes ranging from 20 to 60 with alpha = 0.05.

The number of simulations will be set at 1000 to expedite the analysis. Greater accuracy could be achieved by setting this number higher. This example will still take a few minutes to run because the bootstrap is included in the report.

Setup

This section presents the values of each of the parameters needed to run this example. You can make these changes directly on your screen or you can load the template entitled Example9 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**20 to 60 by 20**
Test Type	**T-Test**
Alternative Hypothesis	**Mean<>M0**
Simulations	**1000**
Distribution|H0 (Null Hypothesis)	**E(M0)**
Distribution|H1 (Alt Hypothesis)	**E(M1)**
M0 (Mean|H0)	**10**
M1 (Mean|H1)	**20**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**
Include T-Test Results	**Checked**
Include Wilcoxon & Sign Test	**Checked**
Include Bootstrap Test Results	**Checked**
Include Exponential Test Results	**Checked**
Iterations Tab	
Bootstrap Iterations	**100**

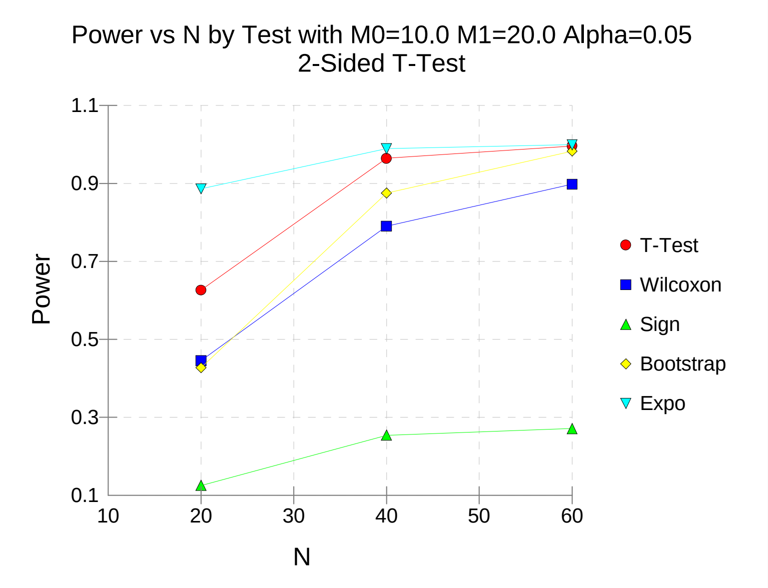
Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Power Comparison for Testing One Mean = Mean0.   Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Expo(M0)
H1 Distribution: Expo(M1)

              H0        H1
            Mean      Mean   Target   T-Test   Wilcoxon     Sign  Bootstrap     Expo
N        (Mean0)   (Mean1)    Alpha    Power      Power    Power      Power    Power
20          10.0      20.0    0.050    0.626      0.445    0.125      0.427    0.886
40          10.0      20.0    0.050    0.964      0.790    0.254      0.875    0.989
60          10.0      20.0    0.050    0.996      0.898    0.271      0.983    0.999
Number of Monte Carlo Iterations: 1000.   Simulation Run Time: 2.43 minutes.
```

```
Alpha Comparison for Testing One Mean = Mean0.   Hypotheses: H0: Mean1=Mean0; H1: Mean1<>Mean0
H0 Distribution: Expo(M0)
H1 Distribution: Expo(M1)

              H0        H1
            Mean      Mean   Target   T-Test   Wilcoxon     Sign  Bootstrap     Expo
N        (Mean0)   (Mean1)    Alpha    Alpha      Alpha    Alpha      Alpha    Alpha
20          10.0      20.0    0.050    0.094      0.130    0.202      0.074    0.048
40          10.0      20.0    0.050    0.057      0.172    0.342      0.046    0.044
60          10.0      20.0    0.050    0.060      0.268    0.489      0.049    0.049
Number of Monte Carlo Iterations: 1000.   Simulation Run Time: 2.43 minutes.
```



As would be expected for exponential data, the exponential test performs the best. The bootstrap test performs nearly as well for larger sample sizes. The other tests fail to achieve the target alpha level. Note that these simulation results will vary from run to run because the samples generated are random. The researcher must now decide which test to use based on her level of confidence in the data being truly exponentially distributed and the size of a sample she can afford to take.

Chapter 415

Non-Inferiority & Superiority Tests for One Mean

Introduction

This module computes power and sample size for non-inferiority and superiority tests in one-sample designs in which the outcome is distributed as a normal random variable. This includes the analysis of the differences between paired values.

The details of sample size calculation for the one-sample design are presented in the Inequality Tests for One Mean (One-Sample or Paired T-Test) chapter and they will not be duplicated here. This chapter only discusses those changes necessary for non-inferiority and superiority tests. Sample size formulas for non-inferiority and superiority tests of a single mean are presented in Chow et al. (2003) page 50.

The *one-sample t-test* is used to test whether a population mean is different from a specific value. When the data are differences between paired values, this test is known as the *paired t-test*. This module also calculates the power of the nonparametric analog of the t-test, the *Wilcoxon test*.

Paired Designs

Paired data may occur because two measurements are made on the same subject or because measurements are made on two subjects that have been matched according to other variables. Hypothesis tests on paired data can be analyzed by considering the difference between the paired items as the response. The distribution of differences is usually symmetric. In fact, the distribution must be symmetric if the individual distributions of the two items are identical. Hence, the paired t-test and the Wilcoxon signed-rank test are appropriate for paired data even when the distributions of the individual items are not normal.

In paired designs, the variable of interest is the difference between two individual measurements. Although the non-inferiority hypothesis refers to the difference between two individual means, the actual values of those means are not needed. All that is needed is their difference.

The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests and their power and sample size could be calculated using the One-Sample T-Test procedure. However, at the urging of our users, we have developed this module which provides the input and output options that are convenient for non-inferiority tests. This section will review the specifics of non-inferiority and superiority testing.

Remember that in the usual t-test setting, the null (H0) and alternative (H1) hypotheses for one-sided tests are defined as

$$\mathrm{H}_0\!:\mu_X \le A \text{ versus } \mathrm{H}_1\!:\mu_X > A$$

Rejecting H0 implies that the mean is larger than the value *A*. This test is called an *upper-tail test* because H0 is rejected in samples in which the sample mean is larger than *A*.

Following is an example of a *lower-tail test*.

$$\mathrm{H}_0\!:\mu_X \ge A \text{ versus } \mathrm{H}_1\!:\mu_X < A$$

Non-inferiority and *superiority* tests are special cases of the above directional tests. It will be convenient to adopt the following specialize notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Population mean*. If the data are paired differences, this is the mean of those differences.
μ_R	Not used	*Reference value*. Usually, this is the mean of a reference population. If the data are paired differences, this is the hypothesized value of the mean difference.
ε	|E|	*Margin of equivalence*. This is a tolerance value that defines the magnitude of difference that is not of practical importance. This may be thought of as the largest difference from the reference value that is considered to be trivial. The absolute value symbols are used to emphasize that this is a magnitude. The sign is determined by the specific design.
δ	D	*True difference*. This is the value of $\mu_T - \mu_R$, the difference between the mean and the reference value, at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their difference is needed for power and sample size calculations.

Non-Inferiority and Superiority Tests

A *non-inferiority test* tests that the mean is not worse than that of the baseline (reference) population by more than a small equivalence margin. The actual direction of the hypothesis depends on the whether higher values of the response are good or bad.

A *superiority test* tests that the mean is better than that of the baseline (reference) population by more than a small equivalence margin. The actual direction of the hypothesis depends on the whether higher values of the response are good or bad.

Case 1: High Values Good, Non-Inferiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the mean of the treatment group is no less than a small amount below the reference value. The value of δ is often set to zero. Equivalent sets of the null and alternative hypotheses are

$\mathrm{H}_0{:}\,\mu_T \le \mu_R - |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T > \mu_R - |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_T - \mu_R \le -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T - \mu_R > -|\varepsilon|$

$\mathrm{H}_0{:}\,\delta \le -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta > -|\varepsilon|$

Case 2: High Values Bad, Non-Inferiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the mean of the treatment group is no more than a small amount above the reference value. The value of δ is often set to zero. Equivalent sets of the null and alternative hypotheses are

$\mathrm{H}_0{:}\,\mu_T \ge \mu_R + |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T < \mu_R + |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_T - \mu_R \ge |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T - \mu_R < |\varepsilon|$

$\mathrm{H}_0{:}\,\delta \ge |\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta < |\varepsilon|$

Case 3: High Values Good, Superiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the mean is greater than the reference value by at least the margin of equivalence. The value of δ must be greater than $|\varepsilon|$. Equivalent sets of the null and alternative hypotheses are

$\mathrm{H}_0{:}\,\mu_T \le \mu_R + |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T > \mu_R + |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_T - \mu_R \le |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T - \mu_R > |\varepsilon|$

$\mathrm{H}_0{:}\,\delta \le |\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta > |\varepsilon|$

Case 4: High Values Bad, Superiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the mean is less than the reference value by at least the margin of equivalence. The value of δ must be less than $-|\varepsilon|$. Equivalent sets of the null and alternative hypotheses are

$$\mathrm{H}_0{:}\,\mu_T \geq \mu_R - |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\,\mu_T < \mu_R - |\varepsilon|$$

$$\mathrm{H}_0{:}\,\mu_T - \mu_R \geq -|\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\,\mu_T - \mu_R < -|\varepsilon|$$

$$\mathrm{H}_0{:}\,\delta \geq -|\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\,\delta < -|\varepsilon|$$

Example

A non-inferiority test example will set the stage for the discussion of the terminology that follows. Suppose that a test is to be conducted to determine if a new cancer treatment adversely affects the mean bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 gm/cm with a standard deviation of 0.000300 gm/cm. Clinicians decide that if the treatment reduces AMBD by more than 5% (0.000115 gm/cm), it poses a significant health threat.

The hypothesis of interest is whether the AMBD in the treated group is greater than 0.002300-0.000115 = 0.002185. The statistical test will be set up so that if the null hypothesis that the AMBD is less than or equal to 0.002185 is rejected, the conclusion will be that the new treatment is non-inferior, at least in terms of AMBD. The value 0.000115 gm/cm is called the *margin of equivalence* or the *margin of non-inferiority*.

Test Statistics

This section describes the test statistics that are available in this procedure.

One-Sample T-Test

The one-sample t-test assumes that the data are a simple random sample from a population of normally-distributed values that all have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the t-test proceeds as follow

$$t_{n-1} = \frac{\overline{X} - D0}{s_{\overline{X}}}$$

where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

$$s_{\overline{X}} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} ,$$

and *D*0 is the value of the mean hypothesized by the null hypothesis.

The significance of the test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the hypothesis is rejected. Otherwise, no conclusion can be reached.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a popular, nonparametric substitute for the t-test. It assumes that the data follow a symmetric distribution. The test is computed using the following steps.

1. Subtract the hypothesized mean, *D*0, from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks *Sp* and the sum of the negative ranks *Sn*. The test statistic, *W*, is the minimum of *Sp* and *Sn*.
3. Compute the mean and standard deviation of *W* using the formulas

$$\mu_{W_n} = \frac{n(n+1)}{4} \quad \text{and} \quad \sigma_{W_n} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}$$

   where *t* represents the number of times the *i*th value occurs.

4. Compute the *z* value using

$$z_W = \frac{W - \mu_{W_n}}{\sigma_{W_n}}$$

The significance of the test statistic is determined by computing the p-value using the standard normal distribution. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, no conclusion can be reached.

Computing the Power

The power is calculated as follows for a directional alternative (one-tailed test) in which *D1* > *D0*. *D1* is the value of the mean at which the power is computed.

1. Find t_α such that $1 - T_{n-1}\left(t_\alpha\right) = \alpha$, where $T_{n-1}\left(t_\alpha\right)$ is the area to the left of *x* under a central-t curve with *n* – 1 degrees of freedom.
2. Calculate $x_a = D0 + t_\alpha \frac{\sigma}{\sqrt{n}}$.
3. Calculate the noncentrality parameter $\lambda = \frac{D1 - D0}{\frac{\sigma}{\sqrt{n}}}$.

4. Calculate $t_a = \dfrac{x_a - D1}{\frac{\sigma}{\sqrt{n}}} + \lambda$.

5. Calculate the power = $1 - T'_{n-1,\lambda}\left(t_a\right)$, where $T'_{n-1,\lambda}\left(x\right)$ is the area to the left of x under a noncentral-t curve with degrees of freedom $n - 1$ and noncentrality parameter λ.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that will be of interest.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power and Beta* or *N*.

Select *N* when you want to determine the sample size needed to achieve a given power and alpha error level.

Select *Power and Beta* when you want to calculate the power.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study. This value must be an integer greater than one. You may enter a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Effect Size – Mean Difference

|E| (Equivalence Margin)

This is the magnitude of the *margin of equivalence*. It is the smallest difference between the mean and the reference value that still results in the conclusion of non-inferiority (or superiority). Note that the sign of this value is assigned depending on the selections for Higher Is and Test Type.

D (True Value)

This is the difference between the mean and the reference value at which the power is computed. For non-inferiority tests, this value is often set to zero, but it can be non-zero as long as the values are consistent with the alternative hypothesis, H1. For superiority tests, this value is non-zero. Again, it must be consistent with the alternative hypothesis, H1.

Effect Size – Standard Deviation

Standard Deviation

This option specifies one or more values of the standard deviation. This must be a positive value. ***PASS*** includes a special module for estimating the standard deviation. This module may be loaded by pressing the *SD* button. Refer to the Standard Deviation Estimator chapter for further details.

Test

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the mean is within the margin of equivalence of being no worse than the reference mean. Select *Superiority* when you want to test whether the mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are usually considered bad. However, if the response variable is income, higher values are generally considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Nonparametric Adjustment

This option makes appropriate sample size adjustments for the Wilcoxon test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Wilcoxon test may be made using the standard *t* test formulations with a simple adjustment to the sample size. The size of the adjustment depends upon the actual distribution of the data. They give sample size adjustment factors for four distributions. These are 1 for the uniform distribution, 2/3 for the double exponential distribution, $9/\pi^2$ for the logistic distribution, and $\pi/3$ for the normal distribution.

The options are as follows:

- **Ignore**

  Do not make a Wilcoxon adjustment. This indicates that you want to analyze a *t* test, not the Wilcoxon test.

- **Uniform**

  Make the Wilcoxon sample size adjustment assuming the uniform distribution. Since the factor is one, this option performs the same as Ignore. It is included for completeness.

- **Double Exponential**

  Make the Wilcoxon sample size adjustment assuming that the data actually follow the double exponential distribution.

- **Logistic**

  Make the Wilcoxon sample size adjustment assuming that the data actually follow the logistic distribution.

- **Normal**

  Make the Wilcoxon sample size adjustment assuming that the data actually follow the normal distribution.

Population Size

This is the number of subjects in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made.

When a finite population size is specified, the standard deviation is reduced according to the formula

$$\sigma_1^2 = \left(1 - \frac{n}{N}\right)\sigma^2$$

where *n* is the sample size, *N* is the population size, σ is the original standard deviation, and σ_1 is the new standard deviation.

The quantity *n/N* is often called the sampling fraction. The quantity $\left(1 - \frac{n}{N}\right)$ is called the *finite population correction factor*.

Example 1 – Power Analysis

Suppose that a test is to be conducted to determine if a new cancer treatment adversely affects the mean bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 gm/cm with a standard deviation of 0.000300 gm/cm. Clinicians decide that if the treatment reduces AMBD by more than 5% (0.000115 gm/cm), it poses a significant health threat. They also want to consider what would happen if the margin of equivalence is set to 2.5% (0.0000575 gm/cm).

Following accepted procedure, the analysis will be a non-inferiority test using the t-test at the 0.025 significance level. Power is to be calculated assuming that the new treatment has no effect on AMBD. Several sample sizes between 20 and 300 will be analyzed. The researchers want to achieve a power of at least 90%. All numbers have been multiplied by 10000 to make the reports and plots easier to read.

Setup

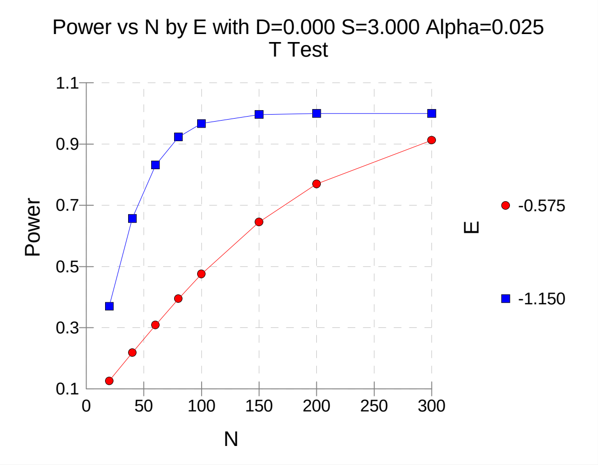
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for One Mean [Differences]** procedure window by clicking on **Means**, then **One Mean**, then **Non-Inferiority & Superiority Tests using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N (Sample Size)	**20 40 60 80 100 150 200 300**
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
S (Standard Deviation)	**3**
Test Type	**Non-Inferiority**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**
Higher is	**Good**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                  Equivalence     Actual  Significance                Standard
                       Margin Difference         Level               Deviation
Power          N          (E)        (D)       (Alpha)        Beta         (S)
0.12601       20       -0.575      0.000       0.02500     0.87399       3.000
0.21844       40       -0.575      0.000       0.02500     0.78156       3.000
0.30873       60       -0.575      0.000       0.02500     0.69127       3.000
0.39493       80       -0.575      0.000       0.02500     0.60507       3.000
0.47532      100       -0.575      0.000       0.02500     0.52468       3.000
0.64517      150       -0.575      0.000       0.02500     0.35483       3.000
0.76959      200       -0.575      0.000       0.02500     0.23041       3.000
0.91262      300       -0.575      0.000       0.02500     0.08738       3.000
0.36990       20       -1.150      0.000       0.02500     0.63010       3.000
0.65705       40       -1.150      0.000       0.02500     0.34295       3.000
0.83164       60       -1.150      0.000       0.02500     0.16836       3.000
0.92317       80       -1.150      0.000       0.02500     0.07683       3.000
0.96682      100       -1.150      0.000       0.02500     0.03318       3.000
0.99658      150       -1.150      0.000       0.02500     0.00342       3.000
0.99970      200       -1.150      0.000       0.02500     0.00030       3.000
1.00000      300       -1.150      0.000       0.02500     0.00000       3.000

Report Definitions
H0 (null hypothesis) is that D <= -|E|, where D = Mean - Reference Value.
H1 (alternative hypothesis) is that D > -|E|.
Power is the probability of rejecting H0 when it is false. It should be close to one.
N is the sample size, the number of subjects in the study.
Alpha is the probability of rejecting H0 when it is true which is the probability of a false positive.
Beta is the probability of accepting H0 when it is false which is the probability of a false negative.
|E| is the magnitude of the margin of equivalence. It is the largest difference that is not of practical significance.
D is actual difference between the mean and the reference value.
Reference Value is a standard value to which the mean is to be compared.
S is the standard deviation of the response. It measures the variability in the population.

Summary Statements
A sample size of 20 achieves 13% power to detect non-inferiority using a one-sided t-test when
the margin of equivalence is -0.575 and the true difference between the mean and the reference
value is 0.000. The data are drawn from a single population with a standard deviation of 3.000.
The significance level (alpha) of the test is 0.02500.
```



The above report shows that for |E| = 1.15, the sample size necessary to obtain 90% power is just under 80. However, if |E| = 0.575, the required sample size is about 300.

Example 2 – Finding the Sample Size

Continuing with Example1, the researchers want to know the exact sample size for each value of |E|.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for One Mean [Differences]** procedure window by clicking on **Means**, then **One Mean**, then **Non-Inferiority & Superiority Tests using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.90**
Alpha	**0.025**
N (Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
S (Standard Deviation)	**3**
Test Type	**Non-Inferiority**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**
Higher is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                   Equivalence      Actual  Significance                Standard
                        Margin  Difference         Level               Deviation
Power          N           (E)         (D)       (Alpha)        Beta         (S)
0.90051      287        -0.575       0.000       0.02500     0.09949       3.000
0.90215       74        -1.150       0.000       0.02500     0.09785       3.000
```

This report shows the exact sample size requirement for each value of |E|.

Example 3 – Validation using Chow

Chow, Shao, Wang (2003) pages 54-55 has an example of a sample size calculation for a non-inferiority trial. Their example obtains a sample size of 8 when D = 0.5, |E| = 0.5, S = 1, Alpha = 0.05, and Beta = 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for One Mean [Differences]** procedure window by clicking on **Means**, then **One Mean**, then **Non-Inferiority & Superiority Tests using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**0.5**
D (True Difference)	**0.5**
S (Standard Deviation)	**1**
Test Type	**Non-Inferiority**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**
Higher is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                     Equivalence       Actual   Significance                 Standard
                          Margin   Difference          Level                Deviation
Power          N             (E)          (D)        (Alpha)         Beta         (S)
0.81502        8          -0.500        0.500        0.05000      0.18498       1.000
```

PASS has also obtained a sample size of 8.

Example 4 – Validation of a Cross-Over Design given in Julious

Julious (2004) page 1953 gives an example of a sample size calculation for a cross-over design. His example obtains a sample size of 87 when D = 0, |E| = 10, S = 28.28427, Alpha = 0.025, and Beta = 0.10. When D is changed to 2, the resulting sample size is 61.

Note that in Julius's example, the population standard deviation is given as 20. Assuming that the correlation between items in a pair is 0, the standard deviation of the difference is calculated to be $S = \sqrt{20^2 + 20^2 - (0)(20)(20)} = 28.284271$. Actually, the value of S probably should be less because the correlation is usually greater than 0 (at least 0.2).

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for One Mean [Differences]** procedure window by clicking on **Means**, then **One Mean**, then **Non-Inferiority & Superiority Tests using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.90**
Alpha	**0.025**
N (Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**10**
D (True Difference)	**0 2**
S (Standard Deviation)	**28.284271**
Test Type	**Non-Inferiority**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**
Higher is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                    Equivalence      Actual  Significance                  Standard
                         Margin  Difference         Level                 Deviation
Power          N            (E)         (D)       (Alpha)         Beta          (S)
0.90332       87        -10.000       0.000       0.02500      0.09668       28.284
0.90323       61        -10.000       2.000       0.02500      0.09677       28.284
```

PASS has also obtained sample sizes of 87 and 61.

Example 5 – Validation of a Cross-Over Design given in Chow, Shao, and Wang

Chow, Shao, and Wang (2004) page 67 give an example of a sample size calculation for a cross-over design. Their example calculates sample sizes of 13 and 14 (13 by formula and 14 from their table) in each sequence (26 or 28 total) when D = -0.1, |E| = 0.2, S = 0.2, Alpha = 0.05, and Beta = 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. You can make these changes directly on your screen or you can load the template entitled Example5 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.80**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**0.2**
D (True Difference)	**-.1**
S (Standard Deviation)	**.2**
Test Type	**Non-Inferiority**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**
Higher is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                      Equivalence     Actual   Significance                  Standard
                           Margin  Difference         Level                  Deviation
Power          N              (E)         (D)       (Alpha)         Beta          (S)
0.81183       27           -0.200      -0.100       0.05000      0.18817        0.200
```

PASS obtained a sample size of 27 which is between the values of 26 and 28 that were obtained by Chow et al.

Chapter 420

Confidence Intervals for One Mean

Introduction

This routine calculates the sample size necessary to achieve a specified distance from the mean to the confidence limit(s) at a stated confidence level for a confidence interval about the mean when the underlying data distribution is normal.

Caution: This procedure assumes that the standard deviation of the future sample will be the same as the standard deviation that is specified. If the standard deviation to be used in the procedure is estimated from a previous sample or represents the population standard deviation, the Confidence Intervals for One Mean with Tolerance Probability procedure should be considered. That procedure controls the probability that the distance from the mean to the confidence limits will be less than or equal to the value specified.

Technical Details

For a single mean from a normal distribution with known variance, a two-sided, 100(1 – α)% confidence interval is calculated by

$$\overline{X} \pm \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}$$

A one-sided 100(1 – α)% upper confidence limit is calculated by

$$\overline{X} + \frac{z_{1-\alpha}\sigma}{\sqrt{n}}$$

Similarly, the one-sided 100(1 – α)% lower confidence limit is

$$\overline{X} - \frac{z_{1-\alpha}\sigma}{\sqrt{n}}$$

For a single mean from a normal distribution with unknown variance, a two-sided, 100(1 – α)% confidence interval is calculated by

$$\overline{X} \pm \frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{\sqrt{n}}$$

A one-sided 100(1 – α)% upper confidence limit is calculated by

$$\overline{X} + \frac{t_{1-\alpha,n-1}\hat{\sigma}}{\sqrt{n}}$$

Similarly, the one-sided 100(1 – α)% lower confidence limit is

$$\overline{X} - \frac{t_{1-\alpha,n-1}\hat{\sigma}}{\sqrt{n}}$$

Each confidence interval is calculated using an estimate of the mean plus and/or minus a quantity that represents the distance from the mean to the edge of the interval. For two-sided confidence intervals, this distance is sometimes called the precision, margin of error, or half-width. We will label this distance, *D*.

The basic equation for determining sample size when D has been specified is

$$D = \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}$$

when the standard deviation is known, and

$$D = \frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{\sqrt{n}}$$

when the standard deviation is unknown. These equations can be solved for any of the unknown quantities in terms of the others. The value $\alpha / 2$ is replaced by α when a one-sided interval is used.

Finite Population Size

The above calculations assume that samples are being drawn from a large (infinite) population. When the population is of finite size (*N*), an adjustment must be made. The adjustment reduces the standard deviation as follows:

$$\sigma_{finite} = \sigma\sqrt{\left(1 - \frac{n}{N}\right)}$$

This new standard deviation replaces the regular standard deviation in the above formulas.

Confidence Level

The confidence level, 1 – α, has the following interpretation. If thousands of samples of n items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean is 1 - α.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters.

Confidence

Confidence Level

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of n items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean is $1 - \alpha$.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90, 0.95* or *0.90 to 0.99 by 0.01*.

Sample Size

N (Sample Size)

Enter one or more values for the sample size. This is the number of individuals selected at random from the population to be in the study.

You can enter a single value or a range of values.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the interval to be used will be a one-sided or a two-sided confidence interval.

Precision

Distance from Mean to Limit(s)

This is the distance from the confidence limit(s) to the mean. For two-sided intervals, it is also known as the precision, half-width, or margin of error.

You can enter a single value or a list of values. The value(s) must be greater than zero.

Standard Deviation

S (Standard Deviation)

Enter a value (or range of values) for the standard deviation. Roughly speaking, this value estimates the average absolute difference between each individual and every other individual. You can use the results of a pilot study, a previous study, or a ball park estimate based on the range (e.g., Range/4) to estimate this parameter.

Know Standard Deviation

Check this box when you want to base your results on the normal distribution. When the box is not checked, calculations are based on the t-distribution. The difference between the two distributions is negligible when the sample sizes are large (>50).

Population

Population Size

This is the number of individuals in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made. This option sets the population size.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Sample Size

Suppose a study is planned in which the researcher wishes to construct a two-sided 95% confidence interval for the mean such that the width of the interval is no wider than 14 units. The confidence level is set at 0.95, but 0.99 is included for comparative purposes. The standard deviation estimate, based on the range of data values, is 28. Instead of examining only the interval half-width of 7, a series of half-widths from 5 to 9 will also be considered.

The goal is to determine the necessary sample size.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95 0.99**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**5 to 9 by 1**
S (Standard Deviation)	**28**
Population Size	**Infinite**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

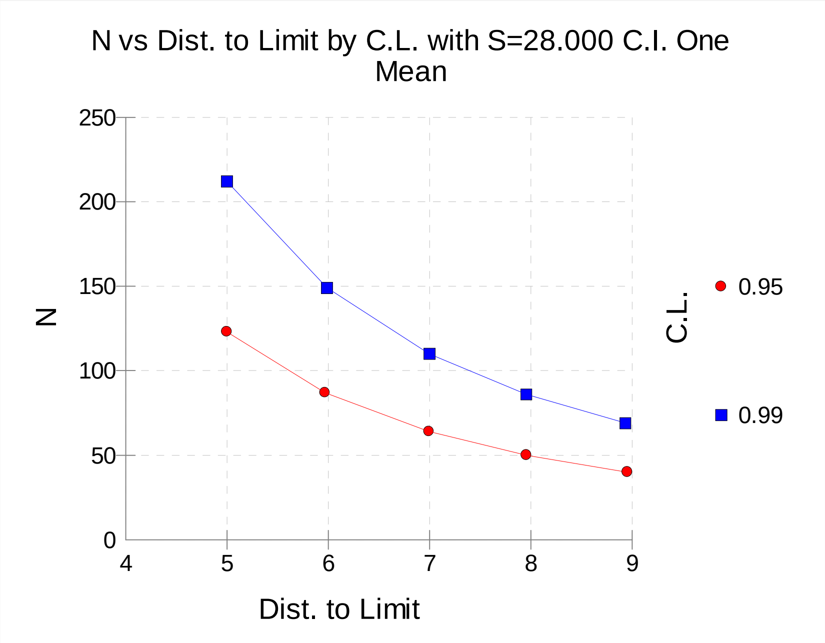
```
Numeric Results for Two-Sided Confidence Intervals with Unknown Standard Deviation

                                   Target      Actual
                      Sample     Distance    Distance      Standard
Confidence              Size    from Mean   from Mean     Deviation
Level                    (N)    to Limits   to Limits           (S)
0.95                     123        5.000       4.998        28.000
0.99                     212        5.000       4.999        28.000
0.95                      87        6.000       5.968        28.000
0.99                     149        6.000       5.986        28.000
0.95                      64        7.000       6.994        28.000
0.99                     110        7.000       6.999        28.000
0.95                      50        8.000       7.958        28.000
0.99                      86        8.000       7.956        28.000
0.95                      40        9.000       8.955        28.000
0.99                      69        9.000       8.933        28.000

References
Hahn, G. J. and Meeker, W.Q. 1991. Statistical Intervals. John Wiley & Sons. New York.
```

Report Definitions
Confidence level is the proportion of confidence intervals (constructed with this same confidence level, sample size, etc.) that would contain the population mean.
N is the size of the sample drawn from the population.
Distance from Mean to Limit is the distance from the confidence limit(s) to the mean. For two-sided intervals, it is also know as the precision, half-width, or margin of error.
Target Distance from Mean to Limit is the value of the distance that is entered into the procedure.
Actual Distance from Mean to Limit is the value of the distance that is obtained from the procedure.
The standard deviation of the population measures the variability in the population.

Summary Statements
A sample size of 123 produces a two-sided 95% confidence interval with a distance from the mean to the limits that is equal to 4.998 when the estimated standard deviation is 28.000.

This report shows the calculated sample size for each of the scenarios.

Plots Section



This plot shows the sample size versus the distance from the mean to the limits (precision) for the two confidence levels.

Example 2 – Validation using Moore and McCabe

Moore and McCabe (1999) page 443 give an example of a sample size calculation for a confidence interval on the mean when the confidence coefficient is 95%, the standard deviation is known to be 3, and the margin of error is 2. The necessary sample size is 9.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**2**
S (Standard Deviation)	**3**
Known Standard Deviation	**Checked**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                                   Target      Actual
                        Sample   Distance    Distance    Standard
Confidence                Size  from Mean   from Mean   Deviation
Level                      (N)  to Limits   to Limits         (S)
0.95                         9      2.000       1.960       3.000
```

PASS also calculated the necessary sample size to be 9.

Example 3 – Validation using Ostle and Malone

Ostle and Malone (1988) page 536 give an example of a sample size calculation for a confidence interval on the mean when the confidence coefficient is 95%, the standard deviation is known to be 7, and the margin of error is 5. The necessary sample size is 8.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**5**
S (Standard Deviation)	**7**
Known Standard Deviation	**Checked**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                                 Target       Actual
                   Sample      Distance     Distance      Standard
Confidence           Size     from Mean    from Mean     Deviation
Level                 (N)     to Limits    to Limits           (S)
0.95                    8         5.000        4.851         7.000
```

PASS also calculated the necessary sample size to be 8.

Chapter 421

Confidence Intervals for One Mean with Tolerance Probability

Introduction

This procedure calculates the sample size necessary to achieve a specified distance from the mean to the confidence limit(s) with a given tolerance probability at a stated confidence level for a confidence interval about a single mean when the underlying data distribution is normal.

Technical Details

For a single mean from a normal distribution with unknown variance, a two-sided, 100(1 – α)% confidence interval is calculated by

$$\bar{X} \pm \frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{\sqrt{n}}$$

A one-sided 100(1 – α)% upper confidence limit is calculated by

$$\bar{X} + \frac{t_{1-\alpha,n-1}\hat{\sigma}}{\sqrt{n}}$$

Similarly, the one-sided 100(1 – α)% lower confidence limit is

$$\bar{X} - \frac{t_{1-\alpha,n-1}\hat{\sigma}}{\sqrt{n}}$$

Each confidence interval is calculated using an estimate of the mean plus and/or minus a quantity that represents the distance from the mean to the edge of the interval. For two-sided confidence intervals, this distance is sometimes called the precision, margin of error, or half-width. We will label this distance, *D*.

The basic equation for determining sample size when D has been specified is

$$D = \frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{\sqrt{n}}$$

Solving for n, we obtain

$$n = \left(\frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{D}\right)^2$$

This equation can be solved for any of the unknown quantities in terms of the others. The value $\alpha/2$ is replaced by α when a one-sided interval is used.

There is an additional subtlety that arises when the standard deviation is to be chosen for estimating sample size. The sample sizes determined from the formula above produce confidence intervals with the specified widths only when the future sample has a sample standard deviation that is no greater than the value specified.

As an example, suppose that 15 individuals are sampled in a pilot study, and a standard deviation estimate of 3.5 is obtained from the sample. The purpose of a later study is to estimate the mean within 10 units. Suppose further that the sample size needed is calculated to be 57 using the formula above with 3.5 as the estimate for the standard deviation. The sample of size 57 is then obtained from the population, but the standard deviation of the 57 individuals turns out to be 3.9 rather than 3.5. The confidence interval is computed and the distance from the mean to the confidence limits is greater than 10 units.

This example illustrates the need for an adjustment to adjust the sample size such that the distance from the mean to the confidence limits will be below the specified value with known probability.

Such an adjustment for situations where a previous sample is used to estimate the standard deviation is derived by Harris, Horvitz, and Mood (1948) and discussed in Zar (1984) and Hahn and Meeker (1991). The adjustment is

$$n = \left(\frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{D}\right)^2 F_{1-\gamma;n-1,m-1}$$

where $1-\gamma$ is the probability that the distance from the mean to the confidence limit(s) will be below the specified value, and m is the sample size in the previous sample that was used to estimate the standard deviation.

The corresponding adjustment when no previous sample is available is discussed in Kupper and Hafner (1989) and Hahn and Meeker (1991). The adjustment in this case is

$$n = \left(\frac{t_{1-\alpha/2,n-1}\hat{\sigma}}{D}\right)^2 \left(\frac{\chi^2_{1-\gamma,n-1}}{n-1}\right)$$

where, again, $1-\gamma$ is the probability that the distance from the mean to the confidence limit(s) will be below the specified value.

Each of these adjustments accounts for the variability in a future estimate of the standard deviation. In the first adjustment formula (Harris, Horvitz, and Mood, 1948), the distribution of the standard deviation is based on the estimate from a previous sample. In the second adjustment formula, the distribution of the standard deviation is based on a specified value that is assumed to be the population standard deviation.

Finite Population Size

The above calculations assume that samples are being drawn from a large (infinite) population. When the population is of finite size (*N*), an adjustment must be made. The adjustment reduces the standard deviation as follows:

$$\sigma_{finite} = \sigma\sqrt{\left(1 - \frac{n}{N}\right)}$$

This new standard deviation replaces the regular standard deviation in the above formulas.

Confidence Level

The confidence level, 1 – α, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean is 1 – α.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters.

Confidence and Tolerance

Confidence Level (1 – Alpha)

The confidence level, 1 – α, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean is 1 – α.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90, 0.95* or *0.90 to 0.99 by 0.01*.

Tolerance Probability

This is the probability that a future interval with sample size N and the specified confidence level will have a distance from the mean to the limit(s) that is less than or equal to the distance specified.

If a tolerance probability is not used, as in the 'Confidence Intervals for One Mean' procedure, the sample size is calculated for the expected distance from the mean to the limit(s), which assumes that the future standard deviation will also be the one specified.

Using a tolerance probability implies that the standard deviation of the future sample will not be known in advance, and therefore, an adjustment is made to the sample size formula to account for the variability in the standard deviation. Use of a tolerance probability is similar to using an upper bound for the standard deviation in the 'Confidence Intervals for One Mean' procedure.

Values between 0 and 1 can be entered. The choice of the tolerance probability depends upon how important it is that the distance from the interval limit(s) to the mean is at most the value specified.

You can enter a range of values such as *0.70 0.80 0.90* or *0.70 to 0.95 by 0.05*.

Sample Size

N (Sample Size)

Enter one or more values for the sample size. This is the number of individuals selected at random from the population to be in the study.

You can enter a single value or a range of values.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the interval to be used will be a one-sided or a two-sided confidence interval.

Precision

Distance from Mean to Limit(s)

This is the distance from the confidence limit(s) to the mean. For two-sided intervals, it is also known as the precision, half-width, or margin of error.

You can enter a single value or a list of values. The value(s) must be greater than zero.

Standard Deviation

Standard Deviation Source

This procedure permits two sources for estimates of the standard deviation:

- **S is a Population Standard Deviation**

  This option should be selected if there is no previous sample that can be used to obtain an estimate of the standard deviation. In this case, the algorithm assumes that future sample obtained will be from a population with standard deviation S.

- **S from a Previous Sample**

  This option should be selected if the estimate of the standard deviation is obtained from a previous random sample from the same distribution as the one to be sampled. The sample size of the previous sample must also be entered under 'Sample Size of Previous Sample'.

Standard Deviation – S is a Population Standard Deviation

S (Standard Deviation)

Enter an estimate of the standard deviation (must be positive). In this case, the algorithm assumes that future samples obtained will be from a population with standard deviation S.

One common method for estimating the standard deviation is the range divided by 4, 5, or 6.

You can enter a range of values such as *1 2 3* or *1 to 10 by 1*.

Press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Standard Deviation – S from a Previous Sample

S (SD Estimated from a Previous Sample)

Enter an estimate of the standard deviation from a previous (or pilot) study. This value must be positive.

A range of values may be entered.

Press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Sample Size of Previous Sample

Enter the sample size that was used to estimate the standard deviation entered in S (SD Estimated from a Previous Sample).

This value is entered only when 'Standard Deviation Source:' is set to 'S from a Previous Sample'.

Population

Population Size

This is the number of individuals in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made. This option sets the population size.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Sample Size

A researcher would like to estimate the mean weight of a population with 95% confidence. It is very important that the mean weight is estimated within 15 grams. Data available from a previous study are used to provide an estimate of the standard deviation. The estimate of the standard deviation is 45.1 grams, from a sample of size 14.

The goal is to determine the sample size necessary to obtain a two-sided confidence interval such that the mean weight is estimated within 15 grams. Tolerance probabilities of 0.70 to 0.95 will be examined.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95**
Tolerance Probability	**0.70 to 0.95 by 0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**15**
Standard Deviation Source	**S from a Previous Sample**
S	**45.1**
Sample Size of Previous Sample	**14**
Population Size	**Infinite**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

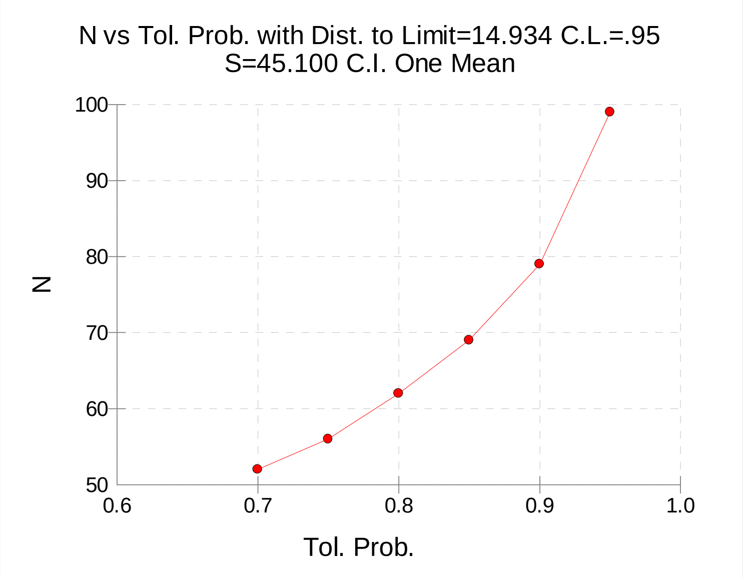
Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals

                                   Target        Actual
                       Sample    Distance      Distance      Standard
Confidence               Size   from Mean     from Mean     Deviation     Tolerance
Level                     (N)   to Limits     to Limits           (S)   Probability
0.95                       52      15.000        14.937        45.100          0.70
0.95                       56      15.000        14.988        45.100          0.75
0.95                       62      15.000        14.907        45.100          0.80
0.95                       69      15.000        14.924        45.100          0.85
0.95                       79      15.000        14.975        45.100          0.90
0.95                       99      15.000        14.934        45.100          0.95

Sample size for estimate of S from previous sample = 10.
```

```
References
Hahn, G. J. and Meeker, W.Q. 1991. Statistical Intervals. John Wiley & Sons. New York.
Zar, J. H. 1984. Biostatistical Analysis. Second Edition. Prentice-Hall. Englewood Cliffs, New Jersey.
Harris, M., Horvitz, D. J., and Mood, A. M. 1948. 'On the Determination of Sample Sizes in Designing
   Experiments', Journal of the American Statistical Association, Volume 43, No. 243, pp. 391-402.

Report Definitions
Confidence level is the proportion of confidence intervals (constructed with this same confidence level,
   sample size, etc.) that would contain the population mean.
N is the size of the sample drawn from the population.
Distance from Mean to Limit is the distance from the confidence limit(s) to the mean. For two-sided intervals,
   it is also know as the precision, half-width, or margin of error.
Target Distance from Mean to Limit is the value of the distance that is entered into the procedure.
Actual Distance from Mean to Limit is the value of the distance that is obtained from the procedure.
The standard deviation of the population measures the variability in the population.
Tolerance Probability is the probability that a future interval with sample size N and corresponding
   confidence level will have a distance from the mean to the limit(s) that is less than or equal to the
   specified distance.

Summary Statements
The probability is 0.70 that a sample size of 52 will produce a two-sided 95% confidence
interval with a distance from the mean to the limits that is less than or equal to 14.937 if
the population standard deviation is estimated to be 45.100 by a previous sample of size 10.
```

This report shows the calculated sample size for each of the scenarios.

Plots Section



This plot shows the sample size versus the tolerance probability.

Example 2 – Validation using Hahn and Meeker

Hahn and Meeker (1991) page 139 give an example of a sample size calculation for a two-sided confidence interval on the mean when the confidence level is 95%, the population standard deviation is assumed to be 2500, the distance from the mean to the limit is 1500, and the tolerance probability is 0.90. The necessary sample size is 19.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95**
Tolerance Probability	**0.90**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**1500**
Standard Deviation Source	**S is a Population Standard Deviation**
S	**2500**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals

                                   Target        Actual
                      Sample     Distance      Distance      Standard
Confidence              Size    from Mean     from Mean     Deviation     Tolerance
Level                    (N)    to Limits     to Limits           (S)   Probability
0.95                      19     1500.000      1447.889      2500.000          0.90
```

PASS also calculated the necessary sample size to be 19.

Example 3 – Validation using Zar

Zar (1984) pages 109-110 give an example of a sample size calculation for a two-sided confidence interval on the mean when the confidence level is 95%, the standard deviation is estimated to be 4.247211 by a previous sample of size 25, the distance from the mean to the limit is 1.5, and the tolerance probability is 0.90. The necessary sample size is 53.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95**
Tolerance Probability	**0.90**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**1.5**
Standard Deviation Source	**S from a Previous Sample**
S	**4.247211**
Sample Size of Previous Sample	**25**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals

                                 Target       Actual
                    Sample     Distance     Distance     Standard
Confidence            Size    from Mean    from Mean    Deviation    Tolerance
Level                  (N)    to Limits    to Limits          (S)  Probability
0.95                    53        1.500        1.489        4.247         0.90
```

PASS also calculated the necessary sample size to be 53.

Example 4 – Validation using Harris, Horvitz, and Mood

Harris, Horvitz, and Mood (1948) pages 392-393 give an example of a sample size calculation for a two-sided confidence interval on the mean when the confidence level is 99%, the standard deviation is estimated to be 3 by a previous sample of size 9, the distance from the mean to the limit is 2, and the tolerance probability is 0.95. The necessary sample size is 49.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for One Mean with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **One Mean with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.99**
Tolerance Probability	**0.90**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**2**
Standard Deviation Source	**S from a Previous Sample**
S	**3**
Sample Size of Previous Sample	**9**
Population Size	**Infinite**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals

                                Target        Actual
                   Sample     Distance      Distance      Standard
Confidence           Size    from Mean     from Mean     Deviation      Tolerance
Level                 (N)    to Limits     to Limits           (S)    Probability
0.99                   49        2.000         1.999         3.000           0.95
```

PASS also calculated the necessary sample size to be 49.

Chapter 430

Inequality Tests for Two Means using Differences (Two-Sample T-Test)

Introduction

A common research task is to compare the means of two populations (groups) by taking independent samples from each. This is sometimes referred to as a *parallel-groups* design. This design is used in situations such as the comparison of the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs.

The mean represents the center of the population. If the means are different, then the populations are different. Other parameters of the two populations (such as the variance) can also be considered, but the mean is usually the starting point.

If assumptions about the other features of the two populations are met (such as that they are normally distributed and their variances are equal), the two-sample *t* test can be used to compare the means of random samples drawn from these two populations. If the normality assumption is violated but the distributions are still symmetric, the nonparametric Mann-Whitney *U* test may be used instead.

Test Procedure

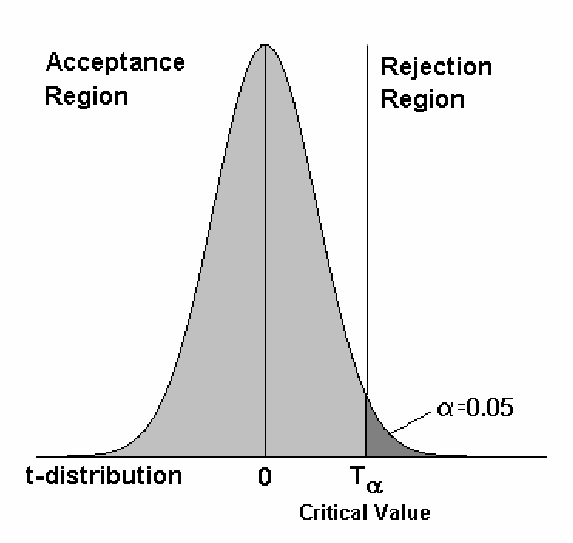
Let the means of populations one and two be μ_1 and μ_2. Let H_0, the *null hypothesis*, represent the hypothesis that the two means are equal. That is, $H_0\!: \mu_1 - \mu_2 = 0$.

The formal steps in conducting a two-sample *t* test and analyzing its power are as follows:

1. **Find the critical value**. Assume that the true difference between the means ($\mu_1 - \mu_2$) is zero. Choose a value T_α so that the probability of rejecting H_0 when H_0 is true is equal to a specified value, α . Using the *t* distribution, select T_α so that $\Pr(T > T_\alpha) = \alpha$.

   Again, select elect T_α so that if the means of the two populations are equal, *t* statistics calculated from two samples drawn from those populations will only exceed T_α exactly $100\alpha\%$ of the time.

   Figure 1 - Find the Critical Value

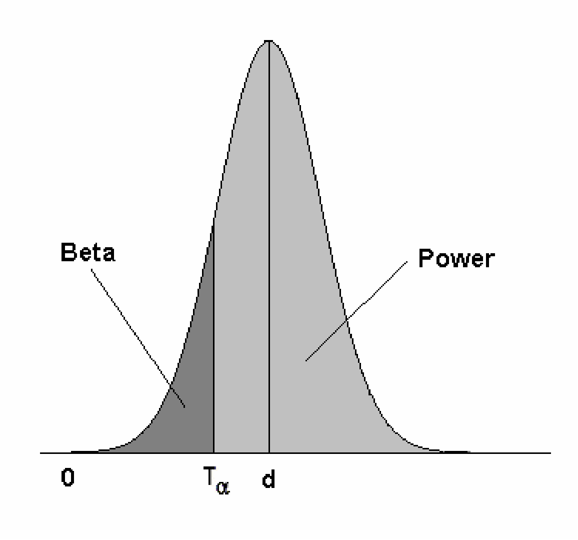
2. **Conduct the experiment**. Select two samples of N_1 and N_2 items from the populations and compute the *t* value. Call this number T_S .

3. **Look for statistical significance**. If $T_S > T_\alpha$ reject the null hypothesis that $\mu_1 - \mu_2 = 0$ in favor of an alternative hypothesis that $\mu_1 - \mu_2 = d > 0$, where $\mu_1 > \mu_2$.

4. **Compute the power**. Now suppose that you want to compute the *power* of this test. First, you must specify an alternative value, *d*, for the difference between the two means so that $\mu_1 = \mu_2 + d$. You now consider a new probability distribution centered at *d* which is called the noncentral-*t* distribution. It appears as a bell-shaped curve as shown below.

   The *power* is the probability of rejecting H_0 when the true difference is *d*. Since we reject H_0 when our computed T_S value is greater that T_α , the power is the area under the noncentral-*t* curve to the right of T_α . The area to the left of T_α represents the probability of a type-II error, or beta, since when the computed T_S value is less than T_α , we do not reject the false H_0 .

Figure 2 - Computing the Power



Notice that in order to compute the power of the test, we must specify the true values of the means. Since we do not know these values, we compute the power at several possible values of *d*. This lets us understand what the power might have been.

Note that we can set the value of alpha (probability of a type-I error). However, we cannot set the value of beta (probability of a type-II error). Beta is computed based on a hypothesized value of *d*. We do not know what the value *d* really is. So we can compute beta for a variety of *d* values, but unless we know the true values of the population means, we do not know the true value of *d*, and hence, we do not know the true value of beta. This is why so much attention is paid to alpha, but so little attention is paid to beta.

Assumptions

The following assumptions are made when using the two-sample *t* test or the Mann-Whitney *U* test. One of the reasons for the popularity of the *t* test is its robustness in the face of assumption violation. However, if an assumption is not met even approximately, the significance levels and the power of the *t* test are unknown. Unfortunately, in practice it often happens that several assumptions are not met. This makes matters even worse! Hence, you should take the appropriate steps to check the assumptions before you make important decisions based on these tests.

Two-Sample T Test Assumptions

The assumptions of the two-sample *t* test are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)
4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired *t* test).
5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Mann-Whitney U Test Assumptions

The assumptions of the Mann-Whitney *U* test for difference in means are:

1. The variable of interest is continuous (not discrete). The measurement scale is at least ordinal.
2. The probability distributions of the two populations are identical, except for location. That is, the variances are equal.
3. The two samples are independent.
4. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Limitations

There are few limitations when using these tests. Sample sizes may range from a few to several hundred. If your data are discrete with at least five unique values, you can often ignore the continuous variable assumption. Perhaps the greatest restriction is that your data come from a random sample of the population. If you do not have a random sample, your significance levels will probably be incorrect.

Technical Details

There are four separate situations each requiring different formulas. Let the means of the two populations be represented by μ_1 and μ_2. The difference between these means will be represented by *d*. Let the standard deviations of the two populations be represented as σ_1 and σ_2.

Case 1 – Standard Deviations Known and Equal

When $\sigma_1 = \sigma_2 = \sigma$ and are known, the power of the *t* test is calculated as follows for a directional alternative (one-tailed test) in which $d > 0$.

1. Find z_α such that $1 - \Phi(z_\alpha) = \alpha$, where $\Phi(x)$ is the area under the standardized normal curve to the left of *x*.
2. Calculate: $\sigma_{\bar{x}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$.
3. Calculate: $z_p = \frac{z_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}}$.
4. Calculate: Power $= 1 - \Phi(z_p)$.

Case 2 – Standard Deviations Known and Unequal

When $\sigma_1 \neq \sigma_2$ and are known, the power is calculated as follows for a directional alternative (one-tailed test) in which *d* > 0.

1. Find z_α such that $1 - \Phi(z_\alpha) = \alpha$, where $\Phi(x)$ is the area under the standardized normal curve to the left of *x*.
2. Calculate: $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$.
3. Calculate: $z_p = \frac{z_\alpha \sigma_{\bar{x}} \text{ - } d}{\sigma_{\bar{x}}}$.
4. Calculate: Power = $1 - \Phi(z_p)$.

Case 3 – Standard Deviations Unknown and Equal

When $\sigma_1 = \sigma_2 = \sigma$ and are unknown, the power of the *t* test is calculated as follows for a directional alternative (one-tailed test) in which *d* > 0.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(t_\alpha)$ is the area under a central-*t* curve to the left of *x* and $df = N_1 + N_2 - 2$.
2. Calculate: $\sigma_{\bar{x}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$.
3. Calculate the noncentrality parameter: $\lambda = \frac{d}{\sigma_{\bar{x}}}$.
4. Calculate: $t_p = \frac{t_\alpha \sigma_{\bar{x}} \text{ - } d}{\sigma_{\bar{x}}} + \lambda$.
5. Calculate: Power = $1 - T'_{df,\lambda}(t_p)$, where $T'_{df,\lambda}(x)$ is the area under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ to the left of *x*.

Case 4 – Standard Deviations Unknown and Unequal

When $\sigma_1 \neq \sigma_2$ and are unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which *d* > 0. Note that in this case, an approximate *t* test is used.

1. Calculate: $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$.

2. Calculate: $f = \dfrac{\sigma_{\bar{x}}^4}{\dfrac{\sigma_1^4}{N_1^2(N_1+1)} + \dfrac{\sigma_2^4}{N_2^2(N_2+1)}} - 2$.

   which is the adjusted degrees of freedom. Often, this is rounded to the next highest integer.

3. Find t_α such that $1 - T_f(t_\alpha) = \alpha$, where $T_f(t_\alpha)$ is the area to the left of *x* under a central-*t* curve with *f* degrees of freedom.

4. Calculate: $\lambda = \dfrac{d}{\sigma_{\bar{x}}}$,1 the noncentrality parameter.

5. Calculate: $t_p = \dfrac{t_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}} + \lambda$.

6. Calculate: Power = $1 - T'_{f,\lambda}(t_p)$, where $T'_{f,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom *f* and noncentrality parameter λ.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Mean1*, *Mean2*, *Sigma1*, *Sigma2*, *Alpha*, *Power and Beta*, *N1*, and *N2*. In most situations, you will select either *Power and Beta* or *N1*.

Select *N1* when you want to calculate the sample size needed to achieve a given power and alpha level.

Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Alpha	Approximate Odds of Rejecting a true null hypothesis
0.01	1 in 100
0.02	1 in 50
0.03	1 in 33
0.04	1 in 25
0.05	1 in 20
0.06	1 in 17
0.07	1 in 14
0.08	1 in 12
0.09	1 in 11
0.10	1 in 10
0.15	1 in 7
0.20	1 in 5
0.25	1 in 4
0.33	1 in 3
0.50	1 in 2

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10*.

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2* = [*R*(*N1*)] where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Effect Size – Means

Mean1 (Mean of Group 1)

This option specifies the mean of the first group. Under the null hypothesis of no difference between groups, the means of both groups are assumed to be equal. Hence, under the null hypothesis, this is also the mean of the second group.

Mean2 (Mean of Group 2)

This option specifies the mean of the second group in the alternative hypothesis. The difference between this value and the value of Mean1 represents the amount that is tested by the *t* test.

Effect Size – Standard Deviations

S1 and S2 (Standard Deviations)

These options specify the values of the standard deviations for each group. When the *S2* is set to *S1*, only *S1* needs to be specified. The value of *S1* will be copied into *S2*.

When these values are not known, you must supply estimates of them. Press the *SD* button to display the Standard Deviation Estimator window. This procedure will help you find appropriate values for the standard deviation.

Known Standard Deviation

This option specifies whether the standard deviations (sigmas) are known or unknown. In almost all experimental situations, sigma is not known. However, since great calculation efficiencies are obtained if we can assume that sigma is known, and since this option has only a small impact on the final result, we usually leave it checked until we are ready for the final results.

When this box is checked, the program makes its calculations assuming that the standard deviations are known. This results in the use of the normal distribution in all probability calculations. Calculations using this option will be much faster than for the unknown sigma case. The results for either case will be close when the sample size is over 30.

When this box is not checked, the program assumes that sigma is not known and will be estimated from the data. This results in probability calculations using the noncentral-*t* distribution. This distribution requires a lot more calculations than does the normal distribution.

The calculation speed comes into play whenever the Find option is set to something besides Beta. In these cases, the program uses a special searching algorithm which requires many iterations. You will note a real difference in calculation speed depending on whether this option is checked.

A reasonable strategy would be to leave this option checked while you are experimenting with the parameters and then leave it unchecked when you are ready for your final results.

Test

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always H_0: *Mean1* = *Mean2*.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **Ha: Mean1 <> Mean2**

  This is the most common selection. It yields the *two-tailed t* test. Use this option when you are testing whether the means are different, but you do not want to specify beforehand which mean is larger.

- **Ha: Mean1 < Mean2**

  This option yields a *one-tailed t* test. Use it when you are only interested in the case in which *Mean2* is greater than *Mean1*.

- **Ha: Mean1 > Mean2**

  This option yields a *one-tailed t* test. Use it when you are only interested in the case in which *Mean2* is less than *Mean1*.

Nonparametric Adjustment (Mann-Whitney Test)

This option lets you make sample size adjustments appropriate for when you are using the Mann-Whitney test rather than the *t* test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Mann-Whitney test may be made using the standard *t* test formulations with a simple adjustment to the sample sizes, *N1* and *N2*. The size of the adjustment depends on the actual distribution of the data. They give sample size adjustment factors for four distributions. These are 1 for uniform, 2/3 for double exponential, $9/\pi^2$ for logistic, and $\pi/3$ for normal.

The options are as follows:

- **Ignore**

  Do not make a Mann-Whitney adjustment. This indicates that you want to analyze a *t* test, not the Mann-Whitney test.

- **Uniform**

  Make the Mann-Whitney sample size adjustment assuming the uniform distribution. Since the factor is one, this option performs the same as Ignore. It is included for completeness.

- **Double Exponential**

  Make the Mann-Whitney sample size adjustment assuming the double exponential distribution.

- **Logistic**

  Make the Mann-Whitney sample size adjustment assuming the logistic distribution.

- **Normal**

  Make the Mann-Whitney sample size adjustment assuming the normal distribution.

Example 1 – Power after a Study

This example will cover the situation in which you are calculating the power of a *t* test after the data have been collected.

A clinical trial was run to compare the effectiveness of two drugs. The ten responses in each group are shown below.

Drug A	Drug B
21	15
20	17
25	17
20	19
23	22
20	12
13	16
18	21
25	20
24	19

These data were run through the ***NCSS*** statistical program with the following results.

```
Descriptive Statistics Section
                                       Standard    Standard    95% LCL     95% UCL
Variable          Count      Mean      Deviation   Error       of Mean     of Mean
Drug A            10         20.9      3.665151    1.159023    18.27811    23.52189
Drug B            10         17.8      3.011091    0.9521905   15.646      19.954

Alternative                            Prob      Decision
Hypothesis                 T Value     Level     (5%)
(Drug A)-(Drug B)<>0       2.0667      0.053460  Accept Ho
```

Notice that the probability level of 0.05346 is not significant. When a test is not significant, its power should be evaluated. The researchers decide to calculate the power using the sample values as estimates for the population values for various sample sizes and for alphas of 0.01 and 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.01 0.05**
N1 (Sample Size Group 1)	**5 10 15 20 25 30 50**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**20.9**
Mean2 (Mean of Group 2)	**17.8**
S1 (Standard Deviation Group 1)	**3.67**
S2 (Standard Deviation Group 2)	**3.01**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Axes/Legend/Grid Tab	
Vertical Range	**Min=0, Max=Data**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

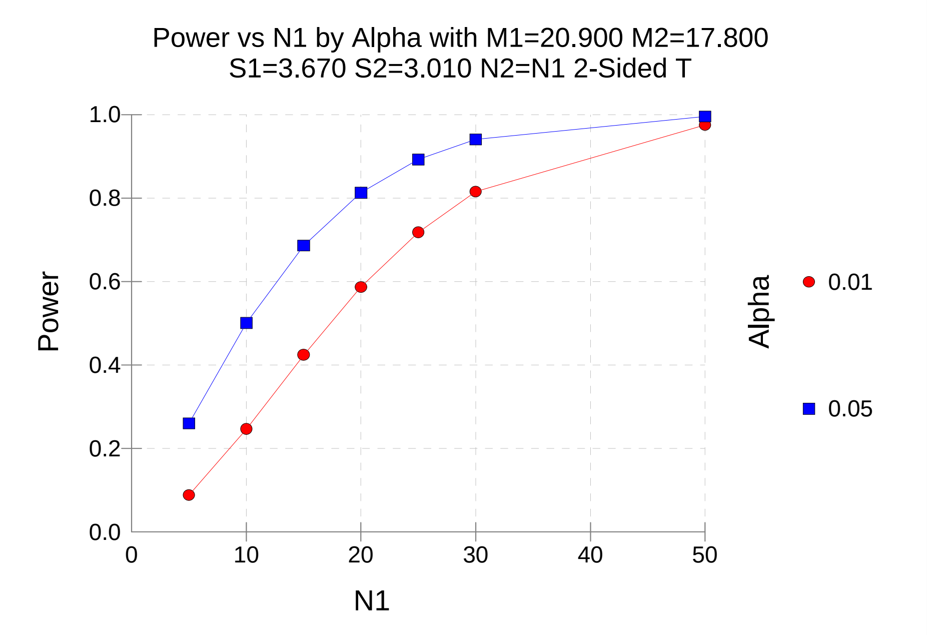
```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and unequal.

                        Allocation
Power      N1     N2       Ratio      Alpha        Beta     Mean1     Mean2        S1        S2
0.08825     5      5        1.00    0.01000     0.91175    20.900    17.800     3.670     3.010
0.24642    10     10        1.00    0.01000     0.75358    20.900    17.800     3.670     3.010
0.42417    15     15        1.00    0.01000     0.57583    20.900    17.800     3.670     3.010
0.58661    20     20        1.00    0.01000     0.41339    20.900    17.800     3.670     3.010
0.71790    25     25        1.00    0.01000     0.28210    20.900    17.800     3.670     3.010
0.81541    30     30        1.00    0.01000     0.18459    20.900    17.800     3.670     3.010
0.97513    50     50        1.00    0.01000     0.02487    20.900    17.800     3.670     3.010
0.26033     5      5        1.00    0.05000     0.73967    20.900    17.800     3.670     3.010
0.50069    10     10        1.00    0.05000     0.49931    20.900    17.800     3.670     3.010
0.68601    15     15        1.00    0.05000     0.31399    20.900    17.800     3.670     3.010
0.81252    20     20        1.00    0.05000     0.18748    20.900    17.800     3.670     3.010
0.89246    25     25        1.00    0.05000     0.10754    20.900    17.800     3.670     3.010
0.94028    30     30        1.00    0.05000     0.05972    20.900    17.800     3.670     3.010
0.99550    50     50        1.00    0.05000     0.00450    20.900    17.800     3.670     3.010
```

Report Definitions
Power is the probability of rejecting a false null hypothesis. Power should be close to one.
N1 and N2 are the number of items sampled from each population. To conserve resources, they should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
Mean1 is the mean of populations 1 and 2 under the null hypothesis of equality.
Mean2 is the mean of population 2 under the alternative hypothesis. The mean of population 1 is unchanged.
S1 and S2 are the population standard deviations. They represent the variability in the populations.

Summary Statements
Group sample sizes of 5 and 5 achieve 16% power to detect a difference of -1.0 between the null hypothesis that both group means are 0.0 and the alternative hypothesis that the mean of group 2 is 1.0 with known group standard deviations of 1.0 and 1.0 and with a significance level (alpha) of 0.01000 using a two-sided two-sample t-test.

This report shows the values of each of the parameters, one scenario per row. At alpha = 0.05 and *N1* = 10, the power was only 0.50. The researchers only had a 50-50 chance of rejecting the null hypothesis in this case.

Plots Section



This plot shows the relationship between alpha and power in this example. Notice that the range of power values over the range of alpha values. Clearly, the sample size should have been doubled to twenty per group in order to achieve a power greater than 0.80.

Example 2 – Finding the Sample Size Necessary to Reject

Continuing with the last example, determine the sample size that the researchers would have needed for the null hypothesis to be rejected at the alpha = 0.01 and 0.05 levels, all other parameters remaining unchanged. They decided to use a beta error level of 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.80**
Alpha	**0.01 0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**20.9**
Mean2 (Mean of Group 2)	**17.8**
S1 (Standard Deviation Group 1)	**3.67**
S2 (Standard Deviation Group 2)	**3.01**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and unequal.

                        Allocation
Power      N1     N2         Ratio      Alpha        Beta      Mean1      Mean2          S1        S2
0.81541    30     30          1.00    0.01000     0.18459     20.900     17.800       3.670     3.010
0.81252    20     20          1.00    0.05000     0.18748     20.900     17.800       3.670     3.010
```

We note that the required sample size is 20 when alpha is 0.05 and 30 when alpha is 0.01. Note that although the power was set at 0.80, the actual power achieved was 0.81. This is due to the fact that sample sizes must be integers, so specified power levels are not met exactly.

Example 3 – Minimum Detectable Difference

The *minimum detectable difference* is the difference between the two means that would be significant if all other parameters are kept at their experimental values. The minimum detectable difference is found by setting Mean1 to zero and solving for Mean2.

Continuing with the previous example, what is the minimum detectable difference when *N1* = *N2* = 10, *alpha* = 0.05, *beta* = 0.20, *S1* = 3.67, and *S2* = 3.01.

Setup

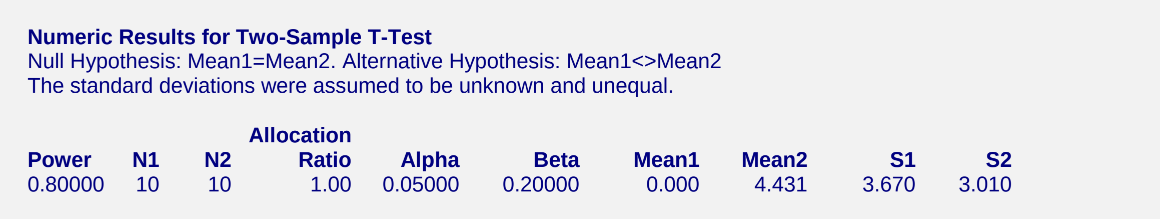
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Mean2 (Search>Mean1)**
Power	**0.80**
Alpha	**0.05**
N1 (Sample Size Group 1)	**10**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	*Ignored since this is the Find setting*
S1 (Standard Deviation Group 1)	**3.67**
S2 (Standard Deviation Group 2)	**3.01**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results



```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and unequal.

                      Allocation
Power     N1    N2       Ratio     Alpha       Beta     Mean1     Mean2        S1        S2
0.80000   10    10        1.00   0.05000    0.20000     0.000     4.431     3.670     3.010
```

The minimum detectable difference for this experiment is 4.431 minutes. If the true population means were this far apart, at a significance level of 0.05 and the power would be 0.80. Hence, the researchers should not have proceeded with the experiment if they thought the true difference was less than 4.431.

Example 4 – Finding the Sample Size

This example will show how the sample size for a new study is determined. A researcher decides to use a *parallel-group design* to study the impact of a new exercise program on body weight. Participants will be divided into two groups: those using and those not using the exercise program. Each participant’s weight loss (or gain) will be measured after three months. How many participants are needed to achieve 90% power at significance levels of 0.01 and 0.05?

Past experiments of this type have had standard deviations in the range of 10 to 15 pounds. The researcher wants to detect a difference of 15 pounds or more.

Although a drop in the mean is hypothesized, two-sided testing will be used because this is the standard method used and the researcher plans on publishing the results.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

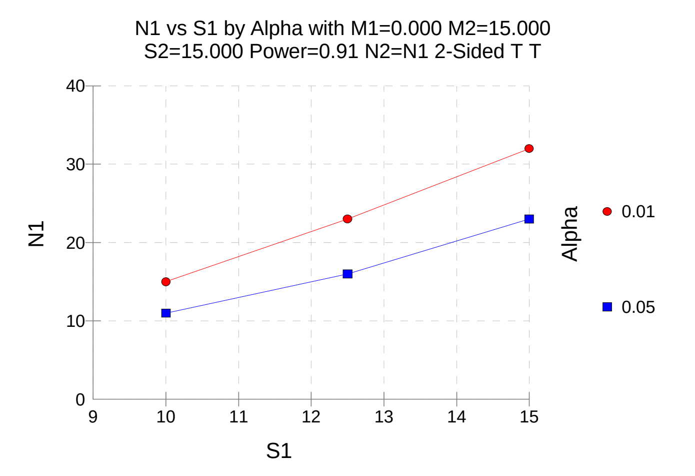
Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.01 0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**15**
S1 (Standard Deviation Group 1)	**10 12.5 15**
S2 (Standard Deviation Group 2)	**S1**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and unequal.

                          Allocation
Power      N1     N2        Ratio     Alpha        Beta      Mean1      Mean2         S1        S2
0.90052    15     15         1.00   0.01000     0.09948      0.000     15.000     10.000    10.000
0.91690    11     11         1.00   0.05000     0.08310      0.000     15.000     10.000    10.000
0.90961    23     23         1.00   0.01000     0.09039      0.000     15.000     12.500    12.500
0.90719    16     16         1.00   0.05000     0.09281      0.000     15.000     12.500    12.500
0.90596    32     32         1.00   0.01000     0.09404      0.000     15.000     15.000    15.000
0.91250    23     23         1.00   0.05000     0.08750      0.000     15.000     15.000    15.000
```



After looking at these reports, the researcher decides to enroll 20 subjects per group and test the hypothesis at the 0.05 significance level. He chooses 20 because it is a little larger than the 16 that are required when the standard deviation is 12.5.

Example 5 – Mann-Whitney Test

The *Mann-Whitney* test is a popular nonparametric analog of the two-sample *t* test. It is recommended when the distribution of the data is not normal. A study by Al-Sunduqchi (1990) showed that sample size and power calculations for the Mann-Whitney test can be made using the standard *t* test results with an adjustment to the sample size.

Suppose that the researcher in Example 4 wants to compare sample size requirements of the *t* test with those of the Mann-Whitney test. To do this, he would use the same values, only this time the Nonparametric Adjustment would be set to a specific distribution. In this example, the double exponential is selected since it requires the largest adjustment of the distributions listed and the actual distribution is not known.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.01 0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**15**
S1 (Standard Deviation Group 1)	**10 12.5 15**
S2 (Standard Deviation Group 2)	**S1**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Double Exponential**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Mann-Whitney Test (Double Exponention Distribution)
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and equal.

                        Allocation
Power      N1     N2         Ratio     Alpha       Beta    Mean1     Mean2       S1        S2
0.90052    10     10          1.00   0.01000    0.09948    0.000    15.000   10.000    10.000
0.91690     7      7          1.00   0.05000    0.08310    0.000    15.000   10.000    10.000
0.90961    15     15          1.00   0.01000    0.09039    0.000    15.000   12.500    12.500
0.90719    10     10          1.00   0.05000    0.09281    0.000    15.000   12.500    12.500
0.90596    21     21          1.00   0.01000    0.09404    0.000    15.000   15.000    15.000
0.91250    15     15          1.00   0.05000    0.08750    0.000    15.000   15.000    15.000
```

Comparing the sample sizes found here with those of the corresponding *t* test found in the last example at the 0.05 significance level, note that there is a reduction in the maximum sample size from 23 to 15. That is, if the Mann-Whitney test is used instead of the *t* test when the actual distribution follows the double exponential distribution, the sample size necessary to achieve 90% power at the 0.05 significance level is reduced from 23 to 15 per group.

Example 6 – Validation of Sample Size using Machin et al.

Machin *et al.* (1997) page 35 present an example in which the mean difference is 5, the common standard deviation is 10, the power is 90%, and the significance level is 0.05. They calculate the per group sample size as 86.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**5**
S1 (Standard Deviation Group 1)	**10**
S2 (Standard Deviation Group 2)	**S1**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and unequal.

                         Allocation
Power      N1     N2          Ratio      Alpha        Beta      Mean1      Mean2         S1         S2
0.90323    86     86           1.00    0.05000     0.09677      0.000      5.000     10.000     10.000
```

Note that the sample size of 86 per group matches Machin's result exactly.

Example 7 – Validation using Zar

Zar (1984) page 136 give an example in which the mean difference is 1, the common standard deviation is 0.7206, the sample sizes are 15 in each group, and the significance level is 0.05. They calculate the power as 0.96.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example7** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**15**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**1**
S1 (Standard Deviation Group 1)	**0.7206**
S2 (Standard Deviation Group 2)	**S1**
Known Standard Deviation	***Not checked***
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<>Mean2
The standard deviations were assumed to be unknown and unequal.

                        Allocation
Power      N1     N2         Ratio      Alpha       Beta      Mean1      Mean2         S1        S2
0.95611    15     15          1.00    0.05000    0.04389      0.000      1.000      0.721     0.721
```

Note that the power of 0.95611 matches Zar's result of 0.96 to the two decimal places given.

Chapter 431

Inequality Tests for Two Means in a Repeated Measures Design

Introduction

This module calculates the power for testing the time-averaged difference (TAD) between two means in a *repeated measures* design. A repeated measures design is one in which subjects are observed repeatedly over time. Measurements may be taken at pre-determined intervals (e.g. weekly or at specified time points following the administration of a particular treatment), or at random times so there are variable intervals between repeated measurements.

Time-averaged difference analysis is often used when the outcome to be measured varies with time. For example, suppose that you want to compare two treatment groups based on the means of a certain outcome such as blood pressure. It is known that a person's blood pressure depends on several instantaneous factors such as amount of sleep, excitement level, mood, exercise, etc. If only a single measurement is taken from each patient then the comparison of mean values from the two groups may be invalid because of the large degree of variation in blood pressure levels among patients. The precision of the experiment is increased by taking multiple measurements from each individual and comparing the time-averaged difference between the two groups. Care must be taken in the analysis because of the correlation that is introduced when several measurements are taken from the same individual. The covariance structure may take on several forms depending on the nature of the experiment and the subjects involved. This procedure allows you to calculate sample sizes using four different covariance patterns: Compound Symmetry, AR(1), Banded(1), and Simple.

This procedure can be used to calculate sample size and power for tests of pairwise contrasts in a mixed models analysis of repeated measures data. Mixed models analysis of repeated measures data is also employed to provide more flexibility in covariance specification and a greater degree of robustness in the presence of missing data, provided that the data can be assumed to be missing at random.

Technical Details

Theory and Notation

For a study with n_1 subjects in group 1 and n_2 subjects in group 2 (for a total of N subjects), each measured m times, the time-averaged difference (d) of a continuous response between two groups can be estimated using the following model:

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}, \quad i = 1, \cdots, N;\ j = 1, \cdots, m,$$

where

y_{ij} is the j^{th} response from subject i,

β_0 is the model intercept,

β_1 is the treatment effect or the time-averaged difference between groups 1 and 2 (i.e. $\beta_1 = d$),

x_i is a binary group assignment variable, which is equal to 1 if the i^{th} subject is in group 1 and equal to 0 if the i^{th} subject is in group 2, and

ε_{ij} is the normal, random error associated with the observation y_{ij}.

Accounting for the relationship between repeated measurements, the model presented above can be written in matrix form as

$$\mathbf{y}_i = \mathbf{X}_i{}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i,$$

where

$\mathbf{y}_i = \begin{pmatrix} y_{i1} & y_{i2} & \cdots & y_{im} \end{pmatrix}'$ is an $m \times 1$ vector of responses from subject i,

$\mathbf{X}_i = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}_{m\times 2}$ if the i^{th} subject is in group 1,

$\mathbf{X}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}_{m\times 2}$ if the i^{th} subject is in group 2,

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ is the vector of model parameters, and

$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{R}_i)$ is the vector of correlated random errors for the observations from subject i,

where

$\mathrm{var}(y_{ij}) = \sigma^2$ is the residual variance for a single observation, and $\mathbf{R}_i$ is the $m \times m$ common correlation matrix for all subjects. The contents of $\mathbf{R}_i$ depend on the assumed within-subject correlation structure.

We can stack the data in a single vector and matrix form as follows:

$$\begin{aligned}\mathbf{y} &= (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_N')' \\ \mathbf{X} &= (\mathbf{X}_1,' \mathbf{X}_2,' \cdots, \mathbf{X}_N')' \\ \boldsymbol{\varepsilon} &= (\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2', \ldots, \boldsymbol{\varepsilon}_N')'\end{aligned}$$

and the model for the N equations can be compressed into one as

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with

$$\begin{aligned}\mathbf{V} &= \mathrm{var}(\mathbf{y}) \\ &= \sigma^2 \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_N \end{pmatrix} \\ &= \sigma^2 \mathbf{R}\end{aligned}$$

as the covariance (or variance - covariance) matrix.

Covariance Pattern

In a repeated measures design with N subjects, each measured m times, observations from a single subject may be correlated, and a pattern for their covariance must be specified. In this case, **V** will have a block-diagonal form:

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V}_N \end{pmatrix},$$

where $\mathbf{V}_i$ are $m \times m$ covariance matrices corresponding to the i^{th} subject. The **0**'s represent $m \times m$ matrices of zeros giving zero covariances for observations on different subjects. This routine allows the specification of four different covariance matrix types: Compound Symmetry, AR(1), Banded(1), and Simple.

Compound Symmetry

A compound symmetry covariance model assumes that all covariances are equal, and all variances on the diagonal are equal. That is

$$\mathbf{V}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \rho & \cdots & \rho \\ \rho & \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \rho & \cdots & 1 \end{pmatrix}_{m \times m}$$

where σ^2 is the residual variance and ρ is the correlation between observations on the same subject.

AR(1)

An AR(1) (autoregressive order 1) covariance model assumes that all variances on the diagonal are equal and that covariances *t* time periods apart are equal to $\sigma^2 \rho^t$. That is

$$\mathbf{V}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{m-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{m-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{m-3} \\ \rho^3 & \rho^2 & \rho & 1 & \dots & \rho^{m-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \rho^{m-4} & \dots & 1 \end{pmatrix}_{m \times m}$$

where σ^2 is the residual variance and ρ is the correlation between observations on the same subject.

Banded(1)

A Banded(1) (banded order 1) covariance model assumes that all variances on the diagonal are equal, covariances for observations one time period apart are equal to $\sigma^2 \rho$, and covariances for measurements greater than one time period apart are equal to zero. That is

$$\mathbf{V}_i = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & 0 & \cdots & 0 \\ \rho & 1 & \rho & 0 & \cdots & 0 \\ 0 & \rho & 1 & \rho & \cdots & 0 \\ 0 & 0 & \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m}$$

where σ^2 is the residual variance and ρ is the correlation between observations on the same subject.

Simple

A simple covariance model assumes that all variances on the diagonal are equal and that all covariances are equal to zero. That is

$$\mathbf{V}_i = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{m \times m}$$

where σ^2 is the residual variance.

Model Estimation

With $\hat{\mathbf{V}} = \hat{\sigma}^2 \hat{\mathbf{R}}$, then estimates of the regression coefficients from the above regression model are given as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\ &= (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \end{aligned}$$

and the variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \operatorname{var}(\hat{\boldsymbol{\beta}}) &= \begin{pmatrix} \operatorname{var}(\hat{\beta}_0) & \operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) & \operatorname{var}(\hat{\beta}_1) \end{pmatrix}, \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \end{aligned}$$

which is estimated by substituting $\hat{\mathbf{V}}$ for **V**.

Hypothesis Test

A two-sided test that the time-averaged difference between the two groups is equal to zero is equivalent to the test of $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Similarly, the upper and lower one-sided tests are $H_0 : \beta_1 \leq 0$ vs. $H_1 : \beta_1 > 0$ and $H_0 : \beta_1 \geq 0$ vs. $H_1 : \beta_1 < 0$, respectively. The test can be carried out using the test statistic

$$z = \frac{\hat{\beta}_1}{\sqrt{\operatorname{var}(\hat{\beta}_1)}} \rightarrow N(0,1).$$

If the standard deviation is unknown and estimated, a *t* test should be used. In practice, this test is often carried out by calculating the average response for each individual and then using a two-sample *t* test. If the data are balanced, the test can also be carried out in ***NCSS*** using Repeated Measures GLM and specifying a comparison such as "Each with First". In the case where the data are not balanced, the test could be carried out using SAS® `PROC MIXED` or SAS® `PROC GLM`. In both cases a `REPEATED` statement should be used, along with a statement such as

```
ESTIMATE 'A-B' treat 1 -1;   or  LSMEANS treat/ PDIFF;
```

Power Calculations

Sample sizes for repeated measures studies are often calculated as if a simple trial with no repeated measures was planned, which results in a higher calculated sample size than would be found if the correlation between repeated measures were taken into consideration. With an idea of the correct covariance structure, and an estimate of the within-patient correlation, you can get a better estimate of the power and sample size necessary to achieve your objectives. If you have no indication of the correct covariance structure for the experiment, then the compound symmetry (program default) is likely to be adequate. If you have no previous estimate of the within-patient correlation, then Brown and Prescott (2006) suggest using a conservative prediction of the correlation, i.e. a higher correlation than anticipated.

For a two-sided test where it is assumed that $d > 0$ (without loss of generality),

$$\begin{aligned}
\text{Power} = 1-\beta &= \Pr(\text{rejecting } H_0 \mid H_1) \\
&= \Pr\left(\left|\frac{\hat{\beta}_1}{\sqrt{\operatorname{var}(\hat{\beta}_1)}}\right| > z_{1-\alpha/2} \mid H_1\right) \\
&\approx \Pr\left(\frac{\hat{\beta}_1}{\sqrt{\operatorname{var}(\hat{\beta}_1)}} > z_{1-\alpha/2} \mid H_1\right) \quad \text{since it is assumed that } d > 0 \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\sqrt{\operatorname{var}(\hat{\beta}_1)}} > z_{1-\alpha/2} - \frac{d}{\sqrt{\operatorname{var}(\hat{\beta}_1)}} \mid H_1\right) \\
&= 1 - \Phi\left(z_{1-\alpha/2} - \frac{d}{\sqrt{\operatorname{var}(\hat{\beta}_1)}}\right),
\end{aligned}$$

where Φ() is the standard normal density function, and α and β are the probabilities of type I and type II error, respectively. For a one-sided test, α is used in place of $\alpha/2$.

Since a t test is usually used to test for a group difference in a case such as this, we should note here that the power calculation using the standard normal distribution represents an approximation to the actual power achieved by the t test. We feel that it might be more appropriate to use the non-central t distribution; however, since the calculation is based on numerous assumptions about the covariance structure that influence the results, it seems unnecessary to worry about the small gain in precision that may occur by using the non-central t distribution. For this reason, along with the fact that this is the published method, we have elected to follow the methods of Brown and Prescott (2006), Liu and Wu (2005), Diggle et al. (1994) and use the standard normal distribution in power and sample size calculations.

Calculating Power for Testing Pairwise Contrasts of Fixed Effects in Mixed Models

Mixed Model Theory and Notation

A linear mixed model incorporates both fixed and random effects. Fixed effects are those effects in the model whose values are assumed constant, or unchanging. Random effects are those effects in the model that are assumed to have arisen from a distribution, resulting in another source of random variation other than residual variation. Brown and Prescott (2006) demonstrates how this methodology may be used to calculate the sample size and power for testing pairwise contrasts of fixed effects in a mixed models analysis of repeated measures data. For an experiment with N subjects, p fixed effect parameters, and q random effect parameters, the general mixed model can be expressed using matrix notation as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \cdots, N$$

where

$\mathbf{y}_i$ is an $n_i \times 1$ vector of responses for subject i,

$\mathbf{X}_i$ is an $n_i \times p$, full-rank design matrix of fixed effects for subject i,

$\boldsymbol{\beta}$ is an $p \times 1$ vector of fixed effects parameters,

$\mathbf{Z}_i$ is an $n_i \times q$ design matrix of the random effects for subject i,

$\mathbf{u}_i$ is a $q \times 1$ vector of random effects for subject i which has means of zero and scaled covariance matrix $\mathbf{G}$,

$\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of errors for subject i with zero mean and scaled covariance $\boldsymbol{\Sigma}_i$.

The covariance of $\mathbf{y}_i$, $\text{var}(\mathbf{y}_i) = \mathbf{V}_i$, can be written as

$$\begin{aligned}\mathbf{V}_i &= \text{var}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}\,\text{var}(\mathbf{u}_i)\mathbf{Z}' + \text{var}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i.\end{aligned}$$

We can stack the data in a single vector and matrix form as follows:

$$\begin{gathered}\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)' \\ \mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N)' \\ \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_N \end{pmatrix} \\ \mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N)' \\ \boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_N)'\end{gathered}$$

and the mixed model for the N equations can be compressed into one as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

with

$$\mathbf{V}=\begin{pmatrix}\mathbf{V}_1 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{V}_N\end{pmatrix}=\mathbf{ZGZ'}+\mathbf{\Sigma}$$

where

$$\mathbf{\Sigma}=\begin{pmatrix}\mathbf{\Sigma}_1 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_N\end{pmatrix}$$

is the covariance (or variance-covariance) matrix.

Mixed Model Estimation

Estimates of the variance components $\mathbf{G}$ and $\mathbf{\Sigma}$ are found using maximum likelihood (ML) or restricted/residual maximum likelihood (REML) methods. From these estimates, $\hat{\mathbf{G}}$ and $\hat{\mathbf{\Sigma}}$, an estimate of $\mathbf{V}$ is obtained as $\hat{\mathbf{V}}=\mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}'+\hat{\mathbf{\Sigma}}$. The fixed effects are then estimated as

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

with the variance of $\hat{\boldsymbol{\beta}}$ estimated as

$$\text{var}(\hat{\boldsymbol{\beta}})=(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}.$$

If $\mathbf{ZGZ'}=\mathbf{0}$ and $\mathbf{\Sigma}=\sigma^2\mathbf{R}$, then these estimation equations are identical to the TAD estimation equations presented earlier, except for the fact that $\boldsymbol{\beta}$ may contain more than two parameters, i.e. a parameter for each fixed effect being modeled. In the TAD model presented above, β_1 represents the difference between two treatment means, *d*. In the mixed model formulation presented here, β_1, β_2, etc. represent individual treatment effects. If there are no random effects, then we can use this routine for TAD to calculate the approximate power for testing pairwise contrasts of fixed effects in mixed models designs.

Brown and Prescott (2006) presents an example on page 228 of an experiment for which the power for testing pairwise contrasts can be calculated using this procedure. To determine the relative efficacy of three treatments in controlling hypertension, patients are assigned to one of the three treatments and blood pressure is measured at four follow-up visits. The study aims to determine the differences in average blood pressure among the three treatments.

Testing Fixed Effects

Significance tests for fixed effects can be done using tests based on the *t* distribution. We can define tests of fixed and random effects as contrasts

$$\mathbf{C}=\mathbf{L}'\hat{\boldsymbol{\beta}}=\mathbf{0},$$

respectively. For example, in a trial containing three treatments A, B, and C, a pairwise comparison of treatments A and C is given by the contrast

$$\mathbf{C}_{AC}=\mathbf{L}'\hat{\boldsymbol{\beta}}=(0\quad 1\quad 0\quad -1)\hat{\boldsymbol{\beta}}=\hat{\beta}_A-\hat{\beta}_C,$$

where the first term in $\boldsymbol{\beta}$ is the intercept term, and the other three terms are the treatment effects. For a single comparison, the test statistic is given by

$$\begin{aligned} t_{df} &= \frac{\mathbf{L}'\hat{\boldsymbol{\beta}}}{\mathrm{SE}(\mathbf{L}'\hat{\boldsymbol{\beta}})} \\ &= \frac{\hat{\beta}_j - \hat{\beta}_h}{\mathrm{SE}(\hat{\beta}_j - \hat{\beta}_h)}, \end{aligned}$$

where *df* is the degrees of freedom, usually determined using the Satterthwaite approximation, and $\hat{\beta}_j$ and $\hat{\beta}_h$ ($j \neq h$) are estimated treatment effects.

Contrasts such as this can be tested in SAS® using the `ESTIMATE` statement or by including the `PDIFF` option in an `LSMEANS` statement. For example, if the variable designating three treatments, A, B, and C, were called "treat", then I could use the following statements in `PROC MIXED` to test for a difference between A and C

```
ESTIMATE 'A-C' treat 1 0 -1;
```

or

```
LSMEANS treat/ PDIFF;
```

The latter statement would produce tests of all pairwise comparisons of the levels of the treatment variable. The former would only test the difference between groups A and C. Of course, these comparison statements must be used in conjunction with appropriate model and class statements (see pages 233-237 of Brown and Prescott (2006) for an example analyzed using SAS® `PROC MIXED`).

Estimates of the correlation (ρ) and the standard deviation (σ) for use in power calculations can be found using SAS® `PROC MIXED`. For a model fit using compound symmetry, σ^2 and ρ can be estimated as the sum of the variance parameters, and the compound symmetry variance parameter divided by the sum of the variance parameters, respectively. For AR(1), Banded(1), and Simple covariance models, σ^2 and ρ can be estimated as the residual variance, and the correlation between adjacent measurements, respectively. Alternatively, the `R` and `RCORR` options may be used within the `REPEATED` statement to display the covariance and correlation matrices, from which the parameter estimates can be determined.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for. When you choose to solve for *n*, the program searches for the lowest sample size that meets the alpha and beta criterion you have specified for each of the terms. The "solve for" parameter is displayed on the vertical axis of the plot.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2* = [*R*(*N1*)] where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Effect Size – Difference to Detect

D1 (Difference|H1)

Enter a value for the treatment difference to be detected. This difference represents a contrast of interest between two treatments in the study. You may enter a single value or a range of values such as *5 10 20* or *5 to 25 by 5*. The items in the list may be separated with commas or blanks.

Effect Size – Repeated Measurements

M (Number of Time Points)

Enter a value for the number of time points (repeated measurements) at which each subject will be observed. You may enter a single value or a range of values such as *3 5 7* or *2 to 8 by 1*. The items in the list may be separated with commas or blanks.

Effect Size – Covariance Structure

Covariance Type

Select the within-subject covariance structure that will be used in the mixed models analysis. The options are:

- **Compound Symmetry**

  All variances on the diagonal of the within-subject variance-covariance matrix are equal to σ^2, and all covariances are equal to $\rho\sigma^2$.

- **AR(1)**

  All variances on the diagonal of the within-subject variance-covariance matrix are equal to σ^2, and the covariance between observations *t* time periods apart is $\rho^t\sigma^2$.

- **Banded(1)**

  All variances on the diagonal of the within-subject variance-covariance matrix are equal to σ^2, and the covariance between observations one time period apart is $\rho\sigma^2$. Covariances between observations more than one time period apart are equal to zero.

- **Simple**

  All variances are equal to σ^2, and all covariances are equal to zero.

Sigma (Std Dev of a Single Observation)

Enter a value for the standard deviation (the square root of the residual variance). This standard deviation is assumed to be equal for the two groups. This parameter is equal to the square root of the sum of the variance parameters when compound symmetry is fit in a mixed models analysis of repeated measures data. This is equal to the square root of the residual variance parameter when an AR(1), Banded(1), or Simple model is fit in a mixed models analysis. You may enter a single value or a range of values such as *5 10 20* or *5 to 25 by 5*. The items in the list may be separated with commas or blanks.

Rho (Autocorrelation)

Enter a value for the correlation between observations on the same subject. When no previous estimate of the within-patient correlation is available, you should use a conservative prediction of the correlation, i.e. a correlation that is higher than anticipated. You may enter a single value or a range of values such as *0.5 0.6 0.7* or *0.4 to 0.9 by 0.1*. The items in the list may be separated with commas or blanks.

Test

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always $H_0: d = 0$.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **One-Sided**

  This option yields a *one-tailed* test. Use it for testing the alternative hypotheses $H_1: d > 0$ or $H_1: d < 0$.

- **Two-Sided**

  This is the most common selection. It yields the *two-tailed* test. Use this option when you are testing whether the means are different, but you do not want to specify beforehand which mean is larger.

Example 1 – Determining Power

Researchers are planning a study of the impact of a new drug on heart rate. They want to evaluate the time-averaged difference in heart rate between subjects who take the new drug, and subjects who take the standard drug. Their experimental protocol calls for a baseline heart rate measurement, followed by administration of a certain level of the drug, followed by three additional measurements 30 minutes apart. They want to be able to detect a 10% difference in heart rate between the two treatments.

Similar studies have found an average heart rate of 93 for individuals taking the standard drug, a standard deviation of 9, and an autocorrelation between adjacent measurements on the same individual of 0.7. The researchers assume that first-order autocorrelation adequately represents the autocorrelation pattern. From a heart rate of 93, a 10% reduction gives 83.7, for a difference of 9.3. The test will be conducted at the 0.05 significance level.

What power does the study achieve over a range of possible sample sizes?

Setup

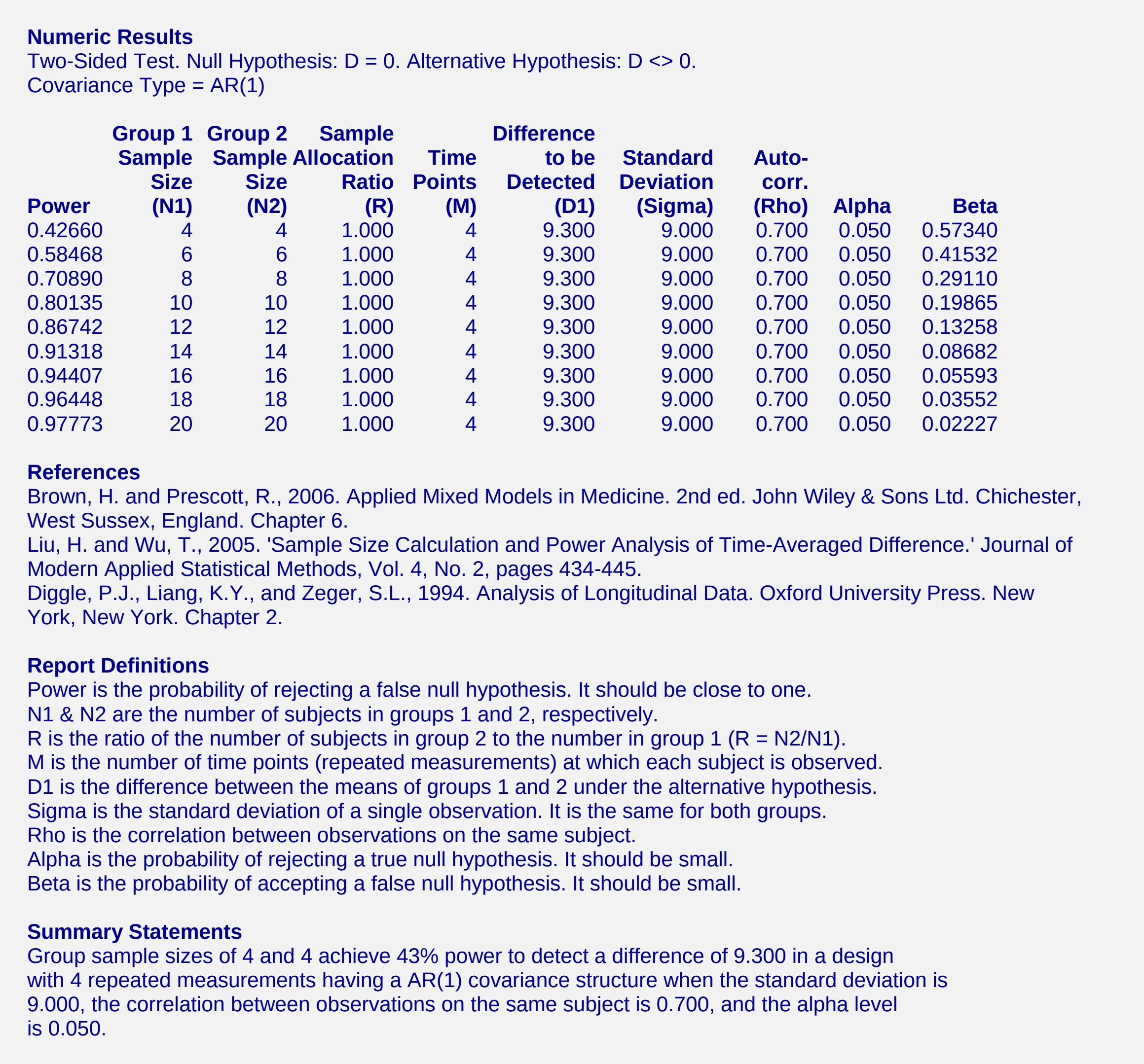
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Repeated Measures Design** procedure window by clicking on **Means**, then **Repeated Measures**, then **Inequality Tests for Two Means in a Repeated Measures Design**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1	**4 to 20 by 2**
N2	**Use R**
R	**1.0**
D1	**9.3**
M	**4**
Covariance Type	**AR(1)**
Sigma	**9**
Rho	**0.7**
Alternative Hypothesis	**Two-Sided**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results



```
Numeric Results
Two-Sided Test. Null Hypothesis: D = 0. Alternative Hypothesis: D <> 0.
Covariance Type = AR(1)

           Group 1  Group 2     Sample         Difference
            Sample   Sample Allocation    Time      to be   Standard   Auto-
              Size     Size      Ratio  Points   Detected  Deviation   corr.
Power         (N1)     (N2)        (R)     (M)       (D1)    (Sigma)   (Rho)  Alpha     Beta
0.42660          4        4      1.000       4      9.300      9.000   0.700  0.050  0.57340
0.58468          6        6      1.000       4      9.300      9.000   0.700  0.050  0.41532
0.70890          8        8      1.000       4      9.300      9.000   0.700  0.050  0.29110
0.80135         10       10      1.000       4      9.300      9.000   0.700  0.050  0.19865
0.86742         12       12      1.000       4      9.300      9.000   0.700  0.050  0.13258
0.91318         14       14      1.000       4      9.300      9.000   0.700  0.050  0.08682
0.94407         16       16      1.000       4      9.300      9.000   0.700  0.050  0.05593
0.96448         18       18      1.000       4      9.300      9.000   0.700  0.050  0.03552
0.97773         20       20      1.000       4      9.300      9.000   0.700  0.050  0.02227

References
Brown, H. and Prescott, R., 2006. Applied Mixed Models in Medicine. 2nd ed. John Wiley & Sons Ltd. Chichester,
West Sussex, England. Chapter 6.
Liu, H. and Wu, T., 2005. 'Sample Size Calculation and Power Analysis of Time-Averaged Difference.' Journal of
Modern Applied Statistical Methods, Vol. 4, No. 2, pages 434-445.
Diggle, P.J., Liang, K.Y., and Zeger, S.L., 1994. Analysis of Longitudinal Data. Oxford University Press. New
York, New York. Chapter 2.

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N1 & N2 are the number of subjects in groups 1 and 2, respectively.
R is the ratio of the number of subjects in group 2 to the number in group 1 (R = N2/N1).
M is the number of time points (repeated measurements) at which each subject is observed.
D1 is the difference between the means of groups 1 and 2 under the alternative hypothesis.
Sigma is the standard deviation of a single observation. It is the same for both groups.
Rho is the correlation between observations on the same subject.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.

Summary Statements
Group sample sizes of 4 and 4 achieve 43% power to detect a difference of 9.300 in a design
with 4 repeated measurements having a AR(1) covariance structure when the standard deviation is
9.000, the correlation between observations on the same subject is 0.700, and the alpha level
is 0.050.
```

This report gives the power for each value of the other parameters.

Power

This is the computed power for detecting the time-averaged difference between the two group means.

Group 1 Sample Size (N1)

The value of *N1* is the number of subjects in group 1.

Group 2 Sample Size (N2)

The value of *N2* is the number of subjects in group 2.

Sample Allocation Ratio (R)

This is the ratio of the number of subjects in group 2 to the number in group 1 (R = N2/N1).

Time Points (M)

This is the number of repeated measurements taken.

Difference to be Detected (D1)

This is the treatment difference that is to be detected.

Standard Deviation (Sigma)

This is the value of σ, the standard deviation or the square root of the residual variance.

Autocorr. (Rho)

This is the correlation between observations from the same subject.

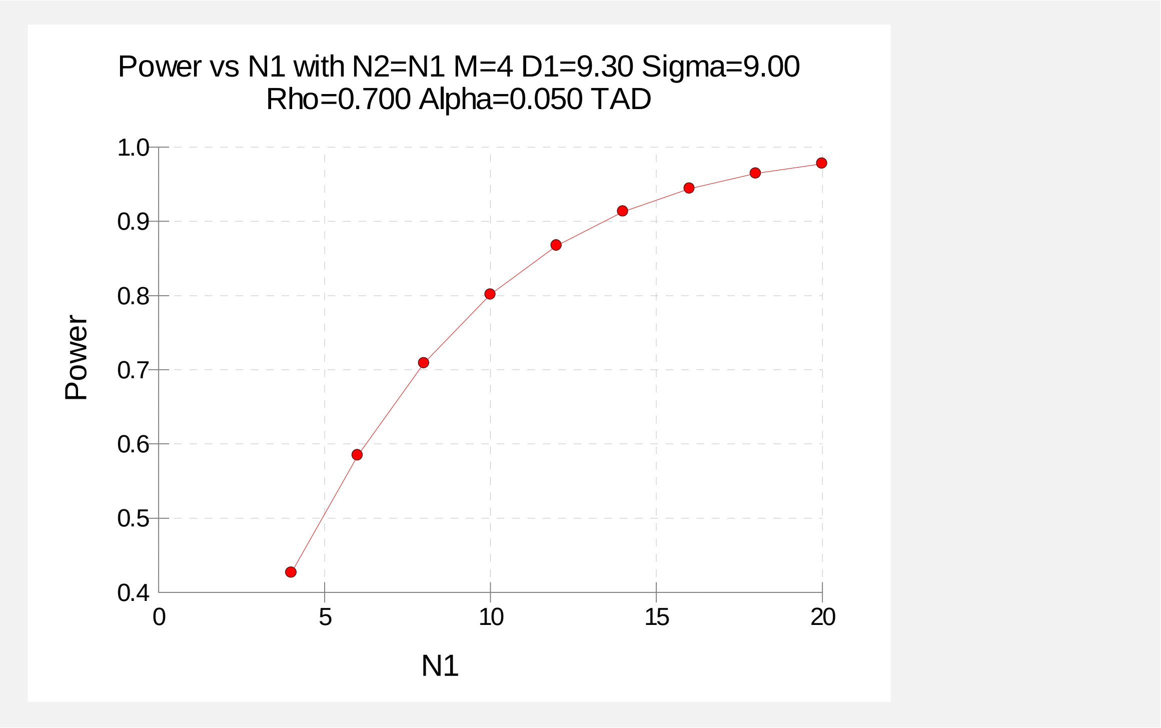
Alpha

Alpha is the significance level of the test.

Beta

Beta is the probability of failing to reject the null hypothesis when the alternative hypothesis is true.

Plots Section



The chart shows the relationship between power and *N1* when the other parameters in the design are held constant.

Example 2 – Finding the Sample Size

Continuing with Example 1, the researchers want to determine the exact sample size necessary to achieve at least 80% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Repeated Measures Design** procedure window by clicking on **Means**, then **Repeated Measures**, then **Inequality Tests for Two Means in a Repeated Measures Design**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find	**N1 (Group 1 Sample Size)**
Power	**0.8**
Alpha	**0.05**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
D1	**9.3**
M	**4**
Covariance Type	**AR(1)**
Sigma	**9**
Rho	**0.7**
Alternative Hypothesis	**Two-Sided**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
         Group 1  Group 2     Sample           Difference
          Sample   Sample Allocation    Time        to be  Standard     Auto-
            Size     Size      Ratio  Points     Detected Deviation      corr.
Power       (N1)     (N2)        (R)     (M)         (D1)   (Sigma)     (Rho)  Alpha      Beta
0.80135       10       10      1.000       4        9.300     9.000     0.700  0.050   0.19865
```

A group sample size of 10 is required to achieve at least 80% power.

Example 3 – Varying the Difference between the Means

Continuing with Examples 1 and 2, the researchers want to evaluate the impact on power of varying the size of the difference between the means for a range of sample sizes from 2 to 8 per group. In the output to follow, we only display the plots. You may want to display the numeric reports as well, but we do not here in order to save space.

Setup

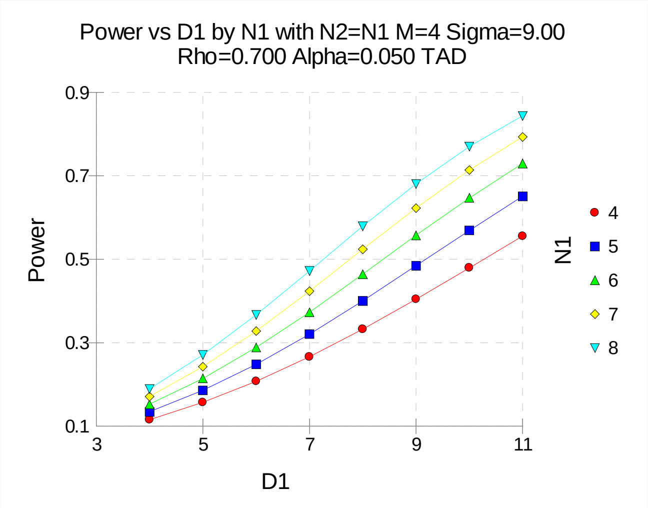
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Repeated Measures Design** procedure window by clicking on **Means**, then **Repeated Measures**, then **Inequality Tests for Two Means in a Repeated Measures Design**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1	**4 to 8 by 1**
N2	**Use R**
R	**1.0**
D1	**4 to 11 by 1**
M	**4**
Covariance Type	**AR(1)**
Sigma	**9**
Rho	**0.7**
Alternative Hypothesis	**Two-Sided**

Output

Click the Run button to perform the calculations and generate the following output.

Plots Section



This chart shows how the power depends on the difference to be detected, *d*, as well as the group sample size, n_1.

Example 4 – Impact of the Number of Repeated Measurements

Continuing with Examples 1 - 3, the researchers want to study the impact on the sample size if they changing the number of measurements made on each individual. Their experimental protocol calls for four measurements that are 30 minutes apart. They want to see the impact of taking twice that many measurements.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Repeated Measures Design** procedure window by clicking on **Means**, then **Repeated Measures**, then **Inequality Tests for Two Means in a Repeated Measures Design**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find	**N1 (Group 1 Sample Size)**
Power	**0.8**
Alpha	**0.05**
N1	*Ignored since this is the Find setting*

Data Tab (continued)

N2	**Use R**
R	**1.0**
D1	**9.3**
M	**4 8**
Covariance Type	**AR(1)**
Sigma	**9**
Rho	**0.7**
Alternative Hypothesis	**Two-Sided**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
          Group 1  Group 2    Sample             Difference
           Sample   Sample Allocation    Time         to be   Standard     Auto-
             Size     Size      Ratio  Points      Detected  Deviation     corr.
Power        (N1)     (N2)        (R)     (M)          (D1)    (Sigma)     (Rho)   Alpha      Beta
0.80135        10       10      1.000       4         9.300      9.000     0.700   0.050   0.19865
0.84737         8        8      1.000       8         9.300      9.000     0.700   0.050   0.15263
```

Doubling the number of repeated measurements per individual decreases the group sample size by 2. This reduction in sample size may not justify the additional four measurements on each subject.

Example 5 – Validation using Diggle et al.

Diggle et al. (1994) page 31 presents an example of calculating the sample size for a TAD study. They calculate the group sample sizes for the cases where d/σ ranges from 0.2 to 0.5, ρ ranges from 0.2 to 0.8, alpha = 0.05, $M = 3$, and power = 0.8. Note that Diggle et al. (1994) uses a one-sided test.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Repeated Measures Design** procedure window by clicking on **Means**, then **Repeated Measures**, then **Inequality Tests for Two Means in a Repeated Measures Design**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find	**N1 (Group 1 Sample Size)**
Power	**0.8**
Alpha	**0.05**
N1	*Ignored since this is the Find setting*

Data Tab (continued)

N2 ....**Use R**
R **1.0**
D1...**0.2 to 0.5 by 0.1**
M **3**
Covariance Type ....**Compound Symmetry**
Sigma **1**
Rho..**0.2 0.5 0.8**
Alternative Hypothesis ....**One-Sided**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

	Group 1	Group 2	Sample		Difference				
	Sample	Sample	Allocation	Time	to be	Standard	Auto-		
	Size	Size	Ratio	Points	Detected	Deviation	corr.		
Power	**(N1)**	**(N2)**	**(R)**	**(M)**	**(D1)**	**(Sigma)**	**(Rho)**	**Alpha**	**Beta**
0.80178	145	145	1.000	3	0.200	1.000	0.200	0.050	0.19822
0.80154	207	207	1.000	3	0.200	1.000	0.500	0.050	0.19846
0.80012	268	268	1.000	3	0.200	1.000	0.800	0.050	0.19988
0.80475	65	65	1.000	3	0.300	1.000	0.200	0.050	0.19525
0.80154	92	92	1.000	3	0.300	1.000	0.500	0.050	0.19846
0.80270	120	120	1.000	3	0.300	1.000	0.800	0.050	0.19730
0.80885	37	37	1.000	3	0.400	1.000	0.200	0.050	0.19115
0.80321	52	52	1.000	3	0.400	1.000	0.500	0.050	0.19679
0.80012	67	67	1.000	3	0.400	1.000	0.800	0.050	0.19988
0.81343	24	24	1.000	3	0.500	1.000	0.200	0.050	0.18657
0.80028	33	33	1.000	3	0.500	1.000	0.500	0.050	0.19972
0.80109	43	43	1.000	3	0.500	1.000	0.800	0.050	0.19891

The sample sizes calculated by ***PASS*** match the results of Diggle et al. (1994) very closely, with slight differences due to rounding. If you calculate the sample sizes by hand, using the formula given in Diggle et al. (1994), page 31, your answers will match those of ***PASS***.

Example 6 – Validation of Sample Size Calculation for Mixed Models Analysis using Brown and Prescott (2006)

Brown and Prescott (2006) pages 268 and 269 presents an example of calculating the sample size for pairwise contrasts in a hypertension trial to by analyzed using mixed models. The analysis of repeated DBP measurements from four post-treatment visits using a compound symmetry covariance pattern resulted in the following covariance matrix for each subject:

$$\mathbf{V}_i = 76\begin{pmatrix} 1 & 0.53 & 0.53 & 0.53 \\ 0.53 & 1 & 0.53 & 0.53 \\ 0.53 & 0.53 & 1 & 0.53 \\ 0.53 & 0.53 & 0.53 & 1 \end{pmatrix}$$

From this matrix they determine that $\rho = 0.53$ and $\sigma^2 = 76\ (\sigma = 8.718)$.

The trial followed several hundred patients given one of three treatments. Brown and Prescott calculate the group sample size to be 31 for a future study involving four post-treatment visits to detect a difference in DBP of 5 mmHg at the 5% significance level with 80% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Repeated Measures Design** procedure window by clicking on **Means**, then **Repeated Measures**, then **Inequality Tests for Two Means in a Repeated Measures Design**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6a** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find	**N1 (Group 1 Sample Size)**
Power	**0.8**
Alpha	**0.05**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
D1	**5**
M	**4**
Covariance Type	**Compound Symmetry**
Sigma	**8.718**
Rho	**0.53**
Alternative Hypothesis	**Two-Sided**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results for M = 4

```
           Group 1  Group 2    Sample         Difference
            Sample   Sample Allocation   Time      to be  Standard    Auto-
              Size     Size      Ratio Points   Detected Deviation    corr.
Power         (N1)     (N2)        (R)    (M)       (D1)   (Sigma)    (Rho)  Alpha      Beta
0.80125         31       31      1.000      4      5.000     8.718    0.530  0.050   0.19875
```

The sample size of 31 calculated by ***PASS*** matches the results of Brown and Prescott (2006) exactly.

Brown and Prescott further calculate the sample size for the case where no account is taken of repeated measurements and the case of 10 repeated measurements. If we change the number of repeated measurements to 1 and 10, we get the following output (Example6b template):

Numeric Results for M = 1, 10

	Group 1	Group 2	Sample		Difference				
	Sample	Sample	Allocation	Time	to be	Standard	Auto-		
	Size	Size	Ratio	Points	Detected	Deviation	corr.		
Power	(N1)	(N2)	(R)	(M)	(D1)	(Sigma)	(Rho)	Alpha	Beta
0.80226	48	48	1.000	1	5.000	8.718	0.530	0.050	0.19774
0.80651	28	28	1.000	10	5.000	8.718	0.530	0.050	0.19349

In both cases, the results of ***PASS*** match those of Brown and Prescott (2006) exactly.

Chapter 435

Inequality Tests for Two Exponential Means

Introduction

This program module designs studies for testing hypotheses about the means of two exponential distributions. Such a test is used when you want to make a comparison between two groups that both follow the exponential distribution. The responses from the samples are assumed to be continuous, positive numbers such as lifetime.

We adopt the basic methodology outlined in the books by Bain and Engelhardt (1991) and Desu and Raghavarao (1990).

Technical Details

The test procedure described here makes the assumption that lifetimes in each group follow an exponential distribution. The densities of the two exponential distributions are written as

$$f_i(t) = \frac{1}{\theta_i}\exp\left(-\frac{t}{\theta_i}\right), \quad i = 1,2$$

The parameters θ_i are interpreted as the average failure times, the mean time to failure (MTTF), or the mean time between failures (MTBF) of the two groups. The reliability, or the probability that a unit continues running beyond time *t*, is

$$R_i(t) = e^{-\frac{t}{\theta_i}}$$

Hypothesis Test

The relevant statistical hypothesis is $H_0: \theta_1 / \theta_2 = 1$ versus one of the following alternatives: $H_A: \theta_1 / \theta_2 = \rho > 1$, $H_A: \theta_1 / \theta_2 = \rho < 1$, or $H_A: \theta_1 / \theta_2 = \rho \neq 1$. The test procedure is to reject the null hypothesis H_0 if the ratio of the observed mean lifetimes $\hat{\rho} = \hat{\theta}_1 / \hat{\theta}_2$ is too large or too small. The samples of size n_i are assumed to be drawn without replacement. The experiment is run until all items fail.

If the experiment is curtailed before all $n_1 + n_2$ items fail, the sample size results are based on the number of failures $r_1 + r_2$, not the total number of samples $n_1 + n_2$.

The mean lifetimes are estimated as follows

$$\hat{\theta}_i = \frac{\sum\limits_{\text{over } j} t_{ij}}{r_i}, \quad i = 1,2$$

where t_{ij} is the time that the *j*th item in the *i*th group is tested, whether measured until failure or until the study is completed.

Power and sample size calculations are based on the fact that the estimated lifetime ratio is proportional to the *F* distribution. That is,

$$\frac{\hat{\theta}_1}{\hat{\theta}_2} \sim \frac{\theta_1}{\theta_2} F_{r_1, r_2}$$

which, under the null hypothesis of equality, becomes

$$\frac{\hat{\theta}_1}{\hat{\theta}_2} \sim F_{r_1, r_2}$$

Note that only the actual numbers of failures are used in these distributions. Hence, we assume that the experiment is run until all items fail so that $r_i = n_i$. That is, the sample sizes are the number of failures, not the number of items. Enough units must be sampled to ensure that the stated number of failures occur.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level. Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power or Beta (Beta is Consumer's Risk)

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta (consumer's risk) is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal thetas when in fact they are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal thetas when in fact they are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10*.

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2* = [*R*(*N1*)] where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Effect Size

Theta1 (Group 1 Mean Life)

Enter one or more values for the *mean life* of group 1 under the alternative hypothesis. This value is usually scaled in terms of elapsed time such as hours, days, or years. Of course, all time values must be on the same time scale.

Note that the value of theta may be calculated from the estimated probability of failure using the relationship

$$\mathrm{P}(Failure) = 1 - e^{-time/\theta}$$

so that

$$\theta = \frac{-time}{\ln\left(1 - P(Failure)\right)}$$

Any positive values are valid. You may enter a range of values such as ‘10 20 30’ or ‘100 to 1000 by 100.’

Note that only the ratio of theta1 and theta2 is used in the calculations.

Theta2 (Group 2 Mean Life)

Enter one or more values for the *mean life* of group 2 under the alternative hypothesis. This value is usually scaled in terms of elapsed time such as hours, days, or years. Of course, all time values must be on the same time scale.

Note that the value of theta may be calculated from the estimated probability of failure using the relationship

$$\mathrm{P}(Failure) = 1 - e^{-time/\theta}$$

so that

$$\theta = \frac{-time}{\ln(1 - P(Failure))}$$

Any positive values are valid. You may enter a range of values such as ‘10 20 30’ or ‘100 to 1000 by 100.’

Note that only the ratio of theta1 and theta2 is used in the calculations.

Test

Alternative Hypothesis

Specify the alternative hypothesis of the test. Since the null hypothesis is equality (a difference between theta1 and theta2 of zero), the alternative is all that needs to be specified.

Note that the alternative hypothesis should match the values of Theta1 and Theta2. That is, if you select Ha: Theta1 > Theta, then the value of Theta1 should be greater than the value of Theta2.

Example 1 – Power for Several Sample Sizes

This example will calculate power for several sample sizes of a study designed to compare the average failure time of (supposedly) identical components manufactured by two companies. Management wants the study to be large enough to detect a ratio of mean lifetimes of 1.3 at the 0.05 significance level. The analysts decide to look at sample sizes between 5 and 500.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Exponential Means** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Exponential Data**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**5 20 50 100 200 300 400 500**

Data Tab (continued)

N2 (Sample Size Group 2)...................... **Use R**
R (Sample Allocation Ratio).................... **1.0**
Theta1 (Group 1 Mean Life).................... **1.3**
Theta2 (Group 2 Mean Life).................... **1.0**
Alternative Hypothesis .... **Ha: Theta1 <> Theta2**

Annotated Output

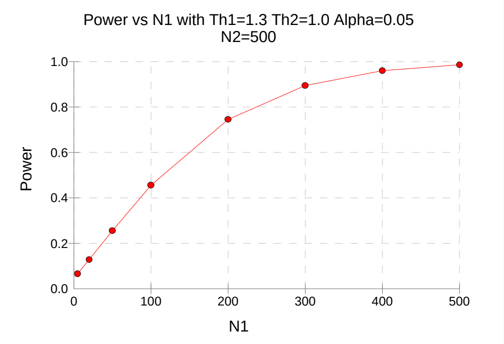
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
H0: Theta1 = Theta2. Ha: Theta1 <> Theta2.
                                   Allocation                                          Theta1/
Power          N1          N2          Ratio         Alpha        Beta  Theta1     Theta2    Theta2
0.06652         5           5        1.00000       0.05000     0.93348     1.3        1.0   1.30000
0.12839        20          20        1.00000       0.05000     0.87161     1.3        1.0   1.30000
0.25602        50          50        1.00000       0.05000     0.74398     1.3        1.0   1.30000
0.45619       100         100        1.00000       0.05000     0.54381     1.3        1.0   1.30000
0.74551       200         200        1.00000       0.05000     0.25449     1.3        1.0   1.30000
0.89447       300         300        1.00000       0.05000     0.10553     1.3        1.0   1.30000
0.95976       400         400        1.00000       0.05000     0.04024     1.3        1.0   1.30000
0.98559       500         500        1.00000       0.05000     0.01441     1.3        1.0   1.30000

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N1 is the number of failures needed in Group 1.
N2 is the number of failures needed in Group 2.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.
Theta1 is the Mean Life in Group 1
Theta2 is the Mean Life in Group 2.

Summary Statements
Samples of size 5 and 5 achieve 7% power to detect a difference between the mean lifetime in
group 1 of 1.3 and the mean lifetime in group 2 of 1.0 at a 0.05000 significance level (alpha)
using a two-sided hypothesis based on the F distribution.
```

This report shows the power for each of the scenarios.

Plots Section



Example 2 – Validation using Manual Calculations

We could not find published results that could be used to validate this procedure. Instead, we will compare the results to those computed using our probability distribution calculator.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Exponential Means** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Exponential Data**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**20**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Theta1 (Group 1 Mean Life)	**1.3**
Theta2 (Group 2 Mean Life)	**1.0**
Alternative Hypothesis	**Ha: Theta1 > Theta2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
H0: Theta1 = Theta2. Ha: Theta1 > Theta2.
                                Allocation                                             Theta1/
Power          N1         N2         Ratio       Alpha       Beta  Theta1    Theta2     Theta2
0.20369        20         20       1.00000     0.05000    0.79631     1.3       1.0    1.30000
```

We will now check these results using manual calculations. First, we find critical value

$$F_{0.95,40,40} = 1.6927972097$$

using the probability calculator. Now, to calculate the power, we find the inverse F of 1.6927972097/1.3 = 1.302152 to be 0.79631, which matches the reported value of Beta.

Chapter 440

Inequality Tests for Two Means (Simulation)

Introduction

This procedure allows you to study the power and sample size of several statistical tests of the null hypothesis that the difference between two means is equal to a specific value versus the alternative hypothesis that it is greater than, less than, or not-equal to that value. Because the mean represents the center of the population, if the means are different, the populations are different. Other attributes of the two populations (such as the shape and spread) might also be compared, but this module focuses on comparisons of the means only.

Measurements are made on individuals that have been randomly assigned to, or randomly chosen from, one of two groups. This is sometimes referred to as a *parallel-groups* design. This design is used in situations such as the comparison of the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs.

The two-sample t-test is commonly used in this situation. When the variances of the two groups are unequal, Welch's t-test is often used. When the data are not normally distributed, the Mann-Whitney (Wilcoxon signed-ranks) U test may be used.

The details of the power analysis of the two-sample t-test using analytic techniques are presented in another ***PASS*** chapter and they won't be duplicated here. This chapter will only consider power analysis using computer simulation.

Technical Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify how the test is carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.

2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's power.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's significance level.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in ***PASS*** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draw the random numbers from this pool. This second method can cut the running time of the simulation by 70%.

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Test Statistics

This section describes the test statistics that are available in this procedure.

Two-Sample T-Test

The two-sample t-test assumes that the data are a simple random sample from a population of normally-distributed values that all have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the t statistic is as follows

$$t_{df} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$\overline{X}_k = \frac{\sum_{i=1}^{N_k} X_{ki}}{N_k}$$

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{\sum_{i=1}^{N_1}\left(X_{1i}-\overline{X}_1\right)^2+\sum_{i=1}^{N_2}\left(X_{2i}-\overline{X}_2\right)^2}{N_1+N_2-2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$$

$$df = N_1+N_2-2$$

The significance of the test statistic is determined by computing the p-value based on the t distribution with degrees of freedom *df*. If this p-value is less than a specified level (often 0.05), the null hypothesis is rejected. Otherwise, no conclusion can be reached.

Welch's T-Test

Welch (1938) proposed the following test for use when the two variances cannot be assumed equal.

$$t_f^* = \frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{s^*_{\overline{X}_1-\overline{X}_2}}$$

where

$$s^*_{\overline{X}_1-\overline{X}_2} = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i}-\overline{X}_1\right)^2}{N_1\left(N_1-1\right)}\right)+\left(\frac{\sum_{i=1}^{N_2}\left(X_{2i}-\overline{X}_2\right)^2}{N_2\left(N_2-1\right)}\right)}$$

$$f = \frac{\left(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\left(N_1-1\right)}+\frac{s_2^4}{N_2^2\left(N_2-1\right)}}$$

$$s_1 = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i}-\overline{X}_1\right)^2}{N_1-1}\right)},\ s_2 = \sqrt{\left(\frac{\sum_{i=1}^{N_2}\left(X_{2i}-\overline{X}_2\right)^2}{N_2-1}\right)}$$

Trimmed T-Test assuming Equal Variances

The notion of trimming off a small proportion of possibly outlying observations and using the remaining data to form a t-test was first proposed for one sample by Tukey and McLaughlin (1963). Dixon and Tukey (1968) consider a slight modification of this one sample test, called

Winsorization, which replaces the trimmed data with the nearest remaining value. The two-sample trimmed t-test was proposed by Yuen and Dixon (1973).

Assume that the data values have been sorted from lowest to highest. The *trimmed mean* is defined as

$$\overline{X}_{tg} = \frac{\sum_{k=g+1}^{N-g} X_k}{h}$$

where $h = N - 2g$ and $g = [N(G/100)]$. Here we use $[Z]$ to mean the largest integer smaller than Z with the modification that if G is non-zero, the value of $[N(G/100)]$ is at least one. G is the percent trimming and should usually be less than 25%, often between 5% and 10%. Thus, the g smallest and g largest observation are omitted in the calculation.

To calculate the modified t-test, calculate the *Winsorized mean* and the *Winsorized* sum of squared deviations as follows.

$$\overline{X}_{wg} = \frac{g\left(X_{g+1} + X_{N-g}\right) + \sum_{k=g+1}^{N-g} X_k}{N}$$

$$SSD_{wg} = \frac{g\left(X_{g+1} - \overline{X}_{wg}\right)^2 + g\left(X_{N-g} - \overline{X}_{wg}\right)^2 + \sum_{k=g+1}^{N-g}\left(X_k - \overline{X}_{wg}\right)^2}{N}$$

Using the above definitions, the two-sample trimmed t-test is given by

$$T_{tg} = \frac{\left(\overline{X}_{1tg} - \overline{X}_{2tg}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{SSD_{1wg} + SSD_{2wg}}{h_1 + h_2 - 2}\left(\frac{1}{h_1} + \frac{1}{h_2}\right)}}$$

The distribution of this t statistic is approximately that of a t distribution with degrees of freedom equal to $h_1 + h_2 - 2$. This approximation is often reasonably accurate if both sample sizes are greater than 6.

Trimmed T-Test assuming Unequal Variances

Yuen (1974) combines trimming (see above) with Welch's (1938) test. The resulting trimmed Welch test is resistant to outliers and seems to alleviate some of the problems that occur because of skewness in the underlying distributions. Extending the results from above, the trimmed version of Welch's t-test is given by

$$T_{tg}^{*} = \frac{\left(\overline{X}_{1tg} - \overline{X}_{2tg}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{SSD_{1wg}}{h_1(h_1 - 1)} + \frac{SSD_{2wg}}{h_2(h_2 - 1)}}}$$

with degrees of freedom *f* given by

$$\frac{1}{f}=\frac{c^2}{h_1-1}+\frac{1-c^2}{h_2-1}$$

where

$$c=\frac{\dfrac{SSD_{1wg}}{h_1\left(h_1-1\right)}}{\dfrac{SSD_{1wg}}{h_1\left(h_1-1\right)}+\dfrac{SSD_{2wg}}{h_2\left(h_2-1\right)}}$$

Mann-Whitney U Test

This test is the nonparametric substitute for the equal-variance t-test. Two key assumptions for this test are that the distributions are at least ordinal and that they are identical under H0. This implies that ties (repeated values) are not acceptable. When ties are present, the approximation provided can be used, but know that the theoretic results no longer hold.

The Mann-Whitney test statistic is defined as follows in Gibbons (1985).

$$z=\frac{W_1-\dfrac{N_1(N_1+N_2+1)}{2}+C}{s_W}$$

where

$$W_1=\sum_{k=1}^{N_1}Rank\left(X_{1k}\right)$$

The ranks are determined after combining the two samples. The standard deviation is calculated as

$$s_W=\sqrt{\frac{N_1N_2(N_1+N_2+1)}{12}-\frac{N_1N_2\sum\limits_{i=1}(t_i^3-t_i)}{12(N_1+N_2)(N_1+N_2-1)}}$$

where t_i is the number of observations tied at value one, t_2 is the number of observations tied at some value two, and so forth.

The correction factor, *C*, is 0.5 if the rest of the numerator of *z* is negative or -0.5 otherwise. The value of *z* is then compared to the standard normal distribution.

Standard Deviations

Care must be used when either the null or alternative distribution is not normal. In these cases, the standard deviation is usually not specified directly. For example, you might use a gamma distribution with a shape parameter of 1.5 and a mean of 4 as the null distribution and a gamma distribution with the same shape parameter and a mean of 5 as the alternative distribution. This allows you to compare the two means. However, although the shape parameters are constant, the standard deviations, which are based on both the shape parameter and the mean, are not. Thus the distributions not only have different means, but different standard deviations!

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data and Options tabs. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the chapter entitled Procedure Window.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies whether you want to find *Power* or *N1* from the simulation. Select *Power* when you want to estimate the power of a certain scenario. Select *N1* when you want to determine the sample size needed to achieve a given power and alpha error level. Finding *N1* is very computationally intensive, and so it may take a long time to complete.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of group 1. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2* = [*R*(*N1*)] where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Test

Test Type

Specify which test statistic is to be used in the simulation. Although the t-test is the most commonly used test statistic, it is based on assumptions that may not be viable in many situations. For your data, you may find that one of the other tests is more accurate (actual alpha = target alpha) and more precise (better power).

Alternative Hypothesis

This option specifies the alternative hypothesis, H1. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always H0: Diff = Diff0.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **Difference <> Diff0**

  This is the most common selection. It yields a *two-tailed test*. Use this option when you are testing whether the mean is different from a specified value Diff0, but you do not want to specify beforehand whether it is smaller or larger. Most scientific journals require two-tailed tests.

- **Difference < Diff0**

  This option yields a *one-tailed test*. Use it when you want to test whether the true mean is less than Diff0.

- **Difference > Diff0**

  This option yields a *one-tailed test*. Use it when you want to test whether the true mean is greater than Diff0. Note that this option could be used for a **non-inferiority test**.

Simulations

Simulations

This option specifies the number of iterations, *M*, used in the simulation. As the number of iterations is increased, the accuracy and running time of the simulation will be increased also.

The precision of the simulated power estimates are calculated from the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the ‘Precision’ amount given in the table.

Simulation	Precision	Precision
Size	when	when
M	Power = 0.50	Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

Group 1 (and 2) Distribution|H0

These options specify the distributions of the two groups under the null hypothesis, H0. The difference between the means of these two distributions is the difference that is tested, *Diff0*.

Usually, these two distributions will be identical and *Diff0* = 0. However, if you are planning a non-inferiority test, the means will be different.

All of the distributions are parameterized so that the mean is entered first. For example, if you wanted to specify that the mean of a normally-distributed variable is to be five, you could enter N(5, S) or N(M0, S) here and *M0* = 5 later.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean is entered first.

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)

Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and are not repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Group 1 (and 2) Distribution|H1

These options specify the distributions of the two groups under the alternative hypothesis, H1. The difference between the means of these two distributions is the difference that is assumed to be the true value of the difference. That is, this is the difference at which the power is computed.

Usually, the mean difference is specified by entering *M0* for the mean parameter in the distribution expression for group 1 and *M1* for the mean parameter in the distribution expression for group 2. The mean difference under H1 then becomes the value of *M0– M1*.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean of group 2 under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean, *M1*, is entered first.

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)

```
Uniform=U(M1,Minimum)
Weibull=W(M1,B)
```

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for *M0* in the distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for *M1* in the distribution specifications given above. Although it can be used wherever you want, *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values for each letter using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Reports Tab

The Reports tab contains settings about the format of the output.

Select Output – Numeric Reports

Show Numeric Reports & Plots

These options let you specify whether you want to generate the standard reports and plots.

Show Inc's & 95% C.I.

Checking this option causes an additional line to be printed showing a 95% confidence interval for both the power and actual alpha and half the width of the confidence interval (the increment).

Select Output – Plots

Show Comparative Reports & Plots

These options let you specify whether you want to generate reports and plots that compare the test statistics that are available.

Comparative Report/Plot Options

Include T-Test Results – Include Mann-Whitney-Test Results

These options let you specify whether to include each test statistic in the comparative reports. These options are only used if comparative reports and/or plots are generated.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size, N1, is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Random Numbers

Random Number Pool Size

This is the size of the pool of random values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Trimmed T-Test

Percent Trimmed at Each End

Specify the percent of each end of the sorted data that is to be trimmed (constant *G* above) when using the trimmed means procedures. This percentage is applied to the sample size to determine how many of the lowest and highest data values are to be trimmed by the procedure. For example, if the sample size (N1) is 27 and you specify 10 here, then [27*10/100] = 2 observations will be trimmed at the bottom and the top. For any percentage, at least one observation is trimmed from each end of the sorted dataset.

The range of possible values is 0 to 25.

Example 1 – Power at Various Sample Sizes

Researchers are planning a parallel-group experiment to test whether the difference in response to a certain drug is zero. The researchers will use a two-sided t-test with an alpha level of 0.05. They want to compare the power at sample sizes of 50, 100, and 200 when the shift in the means is 0.6 from drug 1 to drug 2. They assume that the data are normally distributed with a standard deviation of 2. Since this is an exploratory analysis, they set the number of simulation iterations to 2000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**50 100 200**
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Group 1 Dist’n | H0	**N(M0 S)**
Group 2 Dist’n | H0	**N(M0 S)**
Group 1 Dist’n | H1	**N(M0 S)**
Group 2 Dist’n | H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**
S	**2**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

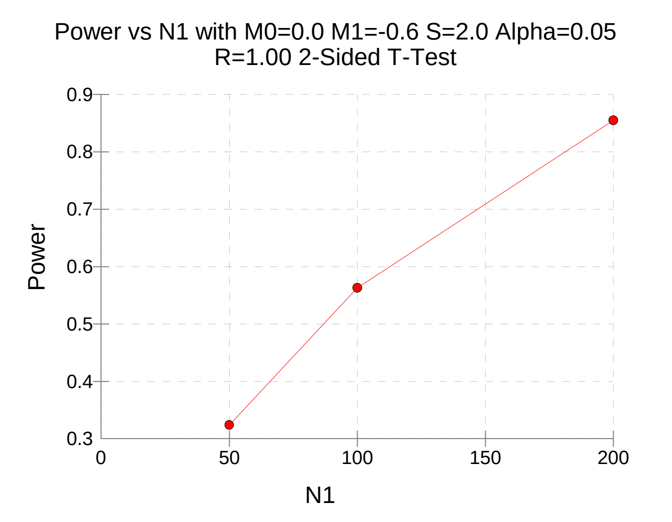
Numeric Results and Plots

```
Numeric Results for Testing Mean Difference = Diff0.    Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist's: Normal(M0 S) & Normal(M0 S)
H1 Dist's: Normal(M0 S) & Normal(M1 S)
Test Statistic: T-Test

                      H0        H1   Target   Actual
Power    N1/N2     Diff0     Diff1    Alpha    Alpha     Beta       M0       M1        S
0.324      50/50     0.0      -0.6    0.050    0.056    0.676      0.0      0.6      2.0
(0.021)  [0.303    0.345]                    (0.010)   [0.045    0.066]

0.563    100/100     0.0      -0.6    0.050    0.047    0.437      0.0      0.6      2.0
(0.022)  [0.541    0.585]                    (0.009)   [0.038    0.056]

0.855    200/200     0.0      -0.6    0.050    0.045    0.145      0.0      0.6      2.0
(0.015)  [0.840    0.870]                    (0.009)   [0.035    0.054]

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 34.78 seconds.
```

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N1 is the size of the sample drawn from population 1.
N2 is the size of the sample drawn from population 2.
Diff0 is the mean difference between (Grp1 - Grp2) assuming the null hypothesis, H0.
Diff1 is the mean difference between (Grp1 - Grp2) assuming the alternative hypothesis, H1.
Target Alpha is the probability of rejecting a true null hypothesis. It is set by the user.
Actual Alpha is the alpha level that was actually achieved by the experiment.
Beta is the probability of accepting a false null hypothesis.
Second Row: (Power Prec.) [95% LCL and UCL Power]   (Alpha Prec.) [95% LCL and UCL Alpha]

Summary Statements
Group sample sizes of 50 and 50 achieve 32% power to detect a difference of -0.6 between the null hypothesis mean difference of 0.0 and the actual mean difference of -0.6 at the 0.050 significance level (alpha) using a two-sided T-Test. These results are based on 2000 Monte Carlo samples from the null distributions: Normal(M0 S) and Normal(M0 S), and the alternative distributions: Normal(M0 S) and Normal(M1 S).

Chart Section



This report shows the estimated power for each scenario. The first row shows the parameter settings and the estimated power and significance level (Actual Alpha).

The second row shows two 95% confidence intervals in brackets: the first for the power and the second for the significance level. Half the width of each confidence interval is given in parentheses as a fundamental measure of the accuracy of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Example 2 – Finding the Sample Size

Continuing with Example1, the researchers want to determine how large a sample is needed to obtain a power of 0.90.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Group 1 Dist’n | H0	**N(M0 S)**
Group 2 Dist’n | H0	**N(M0 S)**
Group 1 Dist’n | H1	**N(M0 S)**
Group 2 Dist’n | H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**
S	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results of Search for N

```
                           H0        H1    Target    Actual
Power        N1/N2       Diff0     Diff1    Alpha     Alpha       Beta         M0          M1         S
0.904      231/231         0.0      -0.6    0.050     0.053      0.097        0.0         0.6       2.0
(0.013)     [0.891      0.916]                       (0.010)    [0.043     0.063]

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 3.00 minutes.
```

The required sample size was 231 which achieved a power of 0.904. To check the accuracy of this simulation, we ran this scenario through the analytic procedure in ***PASS*** which gave the sample size as 234 per group. The simulation answer of 231 was reasonably close.

Example 3 – Comparative Results

Continuing with Example 2, the researchers want to study the characteristics of alternative test statistics. They want to compare the results of all test statistics for N1 = 50, 100, and 200.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**50 100 200**
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Group 1 Dist'n | H0	**N(M0 S)**
Group 2 Dist'n | H0	**N(M0 S)**
Group 1 Dist'n | H1	**N(M0 S)**
Group 2 Dist'n | H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**
S	**2**

Reports Tab

Show Comparative Reports ....**Checked**
Show Comparative Plots.........................**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist's: Normal(M0 S) & Normal(M0 S)
H1 Dist's: Normal(M0 S) & Normal(M1 S)

            H0       H1                          Trim.    Trim.    Mann
           Diff     Diff   Target  T-Test  Welch  T-Test  Welch   Whit'y
N1/N2   (Diff0)  (Diff1)   Alpha   Power   Power   Power  Power   Power      M0      M1      S
50/50       0.0     -0.6   0.050   0.304   0.303   0.283  0.283   0.288     0.0     0.6    2.0
100/100     0.0     -0.6   0.050   0.577   0.577   0.538  0.538   0.544     0.0     0.6    2.0
200/200     0.0     -0.6   0.050   0.859   0.859   0.848  0.848   0.850     0.0     0.6    2.0
Pool Size: 10000. Simulations: 2000. Run Time: 3.66 minutes. Percent Trimmed at each end: 10.

Alpha Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist's: Normal(M0 S) & Normal(M0 S)
H1 Dist's: Normal(M0 S) & Normal(M1 S)

            H0       H1                          Trim.    Trim.    Mann
           Diff     Diff   Target  T-Test  Welch  T-Test  Welch   Whit'y
N1/N2   (Diff0)  (Diff1)   Alpha   Alpha   Alpha   Alpha  Alpha   Alpha      M0      M1      S
50/50       0.0     -0.6   0.050   0.048   0.047   0.048  0.048   0.045     0.0     0.6    2.0
100/100     0.0     -0.6   0.050   0.048   0.048   0.049  0.049   0.048     0.0     0.6    2.0
200/200     0.0     -0.6   0.050   0.054   0.054   0.054  0.054   0.053     0.0     0.6    2.0
Pool Size: 10000. Simulations: 2000. Run Time: 3.66 minutes. Percent Trimmed at each end: 10.
```

These results show that for data that fit the assumptions of the t-test, all five test statistics have accurate alpha values and reasonably close power values. It is interesting to note that the powers of the trimmed procedures, when N1 = 50, are only 7% less than that of the t-test, even though about 20% of the data were trimmed.

Example 4 – Validation using Zar

Zar (1984) page 136 give an example in which the mean difference is 1, the common standard deviation is 0.7206, the sample sizes are 15 in each group, and the significance level is 0.05. They calculate the power to be 0.96.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**15**
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**10000**
Group 1 Dist'n | H0	**N(M0 S)**
Group 2 Dist'n | H0	**N(M0 S)**
Group 1 Dist'n | H1	**N(M0 S)**
Group 2 Dist'n | H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**0.7206**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist's: Normal(M0 S) & Normal(M0 S)
H1 Dist's: Normal(M0 S) & Normal(M1 S)
Test Statistic: T-Test

                        H0        H1   Target   Actual
Power    N1/N2       Diff0     Diff1    Alpha    Alpha      Beta        M0        M1         S
0.956    15/15         0.0      -1.0    0.050    0.045     0.044       0.0       1.0       0.7
(0.004)  [0.952     0.960]                     (0.004)    [0.041    0.049]

Notes:
Pool Size: 20000. Simulations: 10000. Run Time: 10.14 seconds.
```

The power matches the exact value of 0.96.

Example 5 – Non-Inferiority Test

A non-inferiority test is used to show that a new treatment is not significantly worse than the standard (or reference) treatment. The maximum deviation that is ‘not significantly worse’ is called the *margin of equivalence*.

Suppose that the mean diastolic BP of subjects on a certain drug is 96mmHg. If the mean diastolic BP of a new drug is not more than 100mmHg, the drug will be considered non-inferior to the standard drug. The standard deviation among these subjects is 6 mmHg.

The developers of this new drug must design an experiment to test the hypothesis that the mean difference between the two mean BP’s is less than 4. The statistical hypothesis to be tested is

$$H_0: \mu_N - \mu_S \geq 4 \text{ versus } H_1: \mu_N - \mu_S < 4$$

Notice that when the null hypothesis is rejected, the conclusion is that the average difference is less than 4. Following proper procedure, they use a significance level of 0.025 for this one-sided test to keep it comparable to the usual value of 0.05 for a two-sided test. They decide to find the sample size at which the power is 0.90 when the two means are actually equal.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.025**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<Diff0**
Simulations	**2000**
Group 1 Dist'n | H0	**N(M1 S)**
Group 2 Dist'n | H0	**N(M0 S)**
Group 1 Dist'n | H1	**N(M0 S)**
Group 2 Dist'n | H1	**N(M0 S)**
M0 (Mean|H0)	**96**
M1 (Mean|H1)	**100**
S	**6**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                      H0       H1   Target   Actual
Power   N1/N2      Diff0    Diff1    Alpha    Alpha     Beta       M0       M1        S
0.918   49/49        4.0      0.0    0.025    0.024    0.083     96.0    100.0      6.0
```

We see that 49 subjects are required to achieve the desired experimental design.

Example 6 – Selecting a Test Statistic when the Data Contain Outliers

The two-sample t-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracy because the data contain outliers. This example will investigate the impact of outliers on the power and precision of the five test statistics available in ***PASS***.

A mixture of two normal distributions will be used to randomly generate outliers. The mixture will draw 95% of the data from a normal distribution with a mean of 0 and a standard deviation of 1. The other 5% of the data will come from a normal distribution with a mean of 0 and a standard deviation that ranges from 1 to 10.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**20**
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Group 1 Dist’n | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Dist’n | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 1 Dist’n | H1	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Dist’n | H1	**N(M1 S)[95];N(M1 A)[5]**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**1**
A	**1 5 10**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M0 S)[95];Normal(M0 A)[5]
H1 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M1 S)[95];Normal(M1 A)[5]

               H0      H1                                  Trim.   Trim.    Mann
             Diff    Diff  Target  T-Test    Welch  T-Test   Welch  Whit'y
N1/N2     (Diff0)  (Diff1)  Alpha   Power    Power   Power   Power   Power     M0      M1       S      A
20/20         0.0     -1.0  0.050   0.865    0.864   0.835   0.835   0.841    0.0     1.0     1.0    1.0
20/20         0.0     -1.0  0.050   0.638    0.637   0.789   0.787   0.781    0.0     1.0     1.0    5.0
20/20         0.0     -1.0  0.050   0.469    0.463   0.778   0.775   0.776    0.0     1.0     1.0   10.0
Pool Size: 10000. Simulations: 2000. Run Time: 1.77 minutes. Percent Trimmed: 10.
```

```
Alpha Comparison for Testing Mean Difference = Diff0.    Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0

               H0      H1                                  Trim.   Trim.    Mann
             Diff    Diff  Target  T-Test    Welch  T-Test   Welch  Whit'y
N1/N2     (Diff0)  (Diff1)  Alpha   Alpha    Alpha   Alpha   Alpha   Alpha     M0      M1       S      A
20/20         0.0     -1.0  0.050   0.046    0.046   0.045   0.044   0.047    0.0     1.0     1.0    1.0
20/20         0.0     -1.0  0.050   0.040    0.039   0.045   0.044   0.048    0.0     1.0     1.0    5.0
20/20         0.0     -1.0  0.050   0.037    0.034   0.054   0.052   0.061    0.0     1.0     1.0   10.0
Pool Size: 10000. Simulations: 2000. Run Time: 1.77 minutes. Percent Trimmed: 10.
```

The first line gives the results for the standard case in which the two standard deviations (S and A) are equal. Note that in this case, the power of the t-test is a little higher than for the other tests. As the amount of contamination is increased (A equal 5 and then 10), the power of the trimmed tests and the Mann Whitney test remain high, but the power of the t-test falls from 86% to 47%. Also, the value of alpha remains constant for the trimmed and nonparametric tests, but the alpha of the t-test becomes very conservative.

The conclusion this simulation is that if there is a possibility of outliers, you should use either the nonparametric test or the trimmed test.

Example 7 – Selecting a Test Statistic when the Data are Skewed

The two-sample t-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracy when the underlying distributions are skewed. This example will investigate the impact of skewness on the power and precision of the five test statistics available in ***PASS***.

Tukey's lambda distribution will be used because it allows the amount of skewness to be gradually increased.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example7** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**20**
N2 (Sample Size Group 2)	**Use R**
R (Allocation Ratio)	**1.0**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Group 1 Dist’n | H0	**L(M0 S G 0)**
Group 2 Dist’n | H0	**L(M0 S G 0)**
Group 1 Dist’n | H1	**L(M0 S G 0)**
Group 2 Dist’n | H1	**L(M1 S G 0)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**1**
G	**0 0.5 0.9**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist's: Tukey(M0 S G 0) & Tukey(M0 S G 0)
H1 Dist's: Tukey(M0 S G 0) & Tukey(M1 S G 0)

             H0      H1                                Trim.   Trim.   Mann
           Diff    Diff  Target  T-Test    Welch     T-Test   Welch   Whit'y
N1/N2    (Diff0) (Diff1)  Alpha   Power    Power      Power   Power    Power     M0      M1      S      G
20/20       0.0    -1.0   0.050   0.869    0.867      0.833   0.833    0.838    0.0     1.0    1.0    0.0
20/20       0.0    -1.0   0.050   0.880    0.879      0.923   0.922    0.948    0.0     1.0    1.0    0.5
20/20       0.0    -1.0   0.050   0.867    0.866      0.963   0.960    0.993    0.0     1.0    1.0    0.9
Pool Size: 10000. Simulations: 2000. Run Time: 1.85 minutes. Percent Trimmed: 10.

Alpha Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0

             H0      H1                                Trim.   Trim.   Mann
           Diff    Diff  Target  T-Test    Welch     T-Test   Welch   Whit'y
N1/N2    (Diff0) (Diff1)  Alpha   Alpha    Alpha      Alpha   Alpha    Alpha     M0      M1      S      G
20/20       0.0    -1.0   0.050   0.051    0.051      0.043   0.043    0.045    0.0     1.0    1.0    0.0
20/20       0.0    -1.0   0.050   0.039    0.038      0.043   0.041    0.044    0.0     1.0    1.0    0.5
20/20       0.0    -1.0   0.050   0.050    0.049      0.051   0.047    0.054    0.0     1.0    1.0    0.9
Pool Size: 10000. Simulations: 2000. Run Time: 1.85 minutes. Percent Trimmed: 10.
```

The first line gives the results for the standard case in which there is no skewness (G = 0). Note that in this case, the power of the t-test is a little higher than that of the other tests. As the amount of skewness is increased (G equal 0.5 and then 0.9), the power of the trimmed tests and the Mann

Whitney test increases, but the power of the t-test remains about the same. Also, the value of alpha remains constant for all tests.

The conclusion of this simulation is that if there is skewness, you will gain power by using the nonparametric or trimmed test.

Chapter 445

Inequality Tests for Two Means using Ratios (Two-Sample T-Test)

Introduction

This procedure calculates power and sample size for t-tests from a parallel-groups design in which the logarithm of the outcome is a continuous normal random variable. This routine deals with the case in which the statistical hypotheses are expressed in terms of mean ratios instead of mean differences.

The details of testing two treatments using data from a two-group design are given in another chapter, and they will not be repeated here. If the logarithms of the responses can be assumed to follow a normal distribution, hypotheses stated in terms of the ratio can be transformed into hypotheses about the difference. The details of this analysis are given in Julious (2004). They will only be summarized here.

Testing Using Ratios

It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment (group 2) mean.
μ_R	Not used	*Reference mean*. This is the reference (group 1) mean.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only the ratio of these values is needed for power and sample size calculations.

In the two-sided case, the null hypothesis is

$$H_0 \colon \phi = \phi_0$$

and the alternative hypothesis is

$$H_1 \colon \phi \neq \phi_0$$

Log-Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of the ratio of the means.
2. Transform this into hypotheses about a difference by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 are as follows for the null hypothesis.

$$\begin{aligned} &\phi = \phi_0 \\ &\Rightarrow \phi = \left\{ \frac{\mu_T}{\mu_R} \right\} \\ &\Rightarrow \ln(\phi) \neq \left\{ \ln(\mu_T) - \ln(\mu_R) \right\} \end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter can be used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable *X* is the logarithm of the original variable *Y*. That is, *X* = ln(*Y*) and *Y* = exp(*X*). Label the mean and variance of *X* as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of *Y* as μ_Y and σ_Y^2, respectively. If *X* is normally distributed, then *Y* is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be found to be

$$COV_Y = \frac{\sqrt{\mu_Y^2\left(e^{\sigma_X^2}-1\right)}}{\mu_Y}$$
$$= \sqrt{e^{\sigma_X^2}-1}$$

Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2+1\right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2+1\right)}$$

One final note: for parallel-group designs, σ_X^2 equals σ_d^2, the average variance used in the t-test of the logged data.

Thus, the hypotheses can be stated in the original (Y) scale and then the power can be analyzed in the transformed (X) scale.

Power Calculation

As is shown above, the hypotheses can be stated in the original (Y) scale using ratios or the logged (X) scale using differences. In either case, the power and sample size calculations are made using the formulas for testing the difference in two means. These formulas are presented in another chapter and are not duplicated here.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. In most situations, you will select either *Power and Beta* for a power analysis or *N1* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size)

Enter a value (or range of values) for the sample size of group 1 (the reference group). Note that these values are ignored when you are solving for N1. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 (the treatment group) or enter *Use R* to base N2 on the value of N1. You may enter a range of values such as *10 to 100 by 10*.

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when N2 is set to *Use R*.

When used, N2 is calculated from N1 using the formula: N2 = [R(N1)] where [Y] is the next integer greater than or equal to Y. Note that setting R = 1.0 forces N2 = N1.

Effect Size – Ratios

R0 (Ratio Under H0)

This is the value of the ratio of the two means assumed by the null hypothesis, H0. Usually, R0 = 1.0 which implies that the two means are equal. However, you may test other values of R0 as well. Strictly speaking, any positive number is valid, but values near to, or equal to, 1.0 are usually used.

Warning: you cannot use the same value for both R0 and R1.

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Often, a range of values will be tried. For example, you might try the four values:

1.05 1.10 1.15 1.20

Strictly speaking, any positive number is valid. However, numbers between 0.50 and 2.00 are usually used.

Warning: you cannot use the same value for both R0 and R1.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not log) scale. This value must be determined from past experience or from a pilot study. See the discussion above for more details on the definition of the coefficient of variation.

Test

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. Possible selections are:

- **H1: R1 <> R0**

  This is the most common selection. It yields the *two-tailed t-test*. Use this option when you are testing whether the means are different, but you do not want to specify beforehand which mean is larger.

- **H1: R1 < R0**

  This option yields a *one-tailed t-test*. Use it when you are only interested in the case in which *Mean1* is greater than *Mean2*.

- **H1: R1 > R0**

  This option yields a *one-tailed t-test*. Use it when you are only interested in the case in which *Mean1* is less than *Mean2*.

Example 1 – Finding Power

A company has developed a generic drug for treating rheumatism and wants to show that it is better than the standard drug. From previous studies, responses for either treatment are known to follow a lognormal distribution. A parallel-group design will be used and the logged data will be analyzed with a one-sided, two-sample t-test.

Past experience leads the researchers to set the COV to 1.20. The significance level is 0.025. The power will be computed for R1 equal 1.10 and 1.20. Sample sizes between 100 and 900 will be examined in the analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Ratios (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N1 (Sample Size Group 1)	**100 to 900 by 200**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
R0 (Ratio under H0)	**1.0**
R1 (True Ratio)	**1.1 1.2**
COV (Coefficient of Variation)	**1.2**
Alternative Hypothesis	**R1>R0 (One-Sided)**

Annotated Output

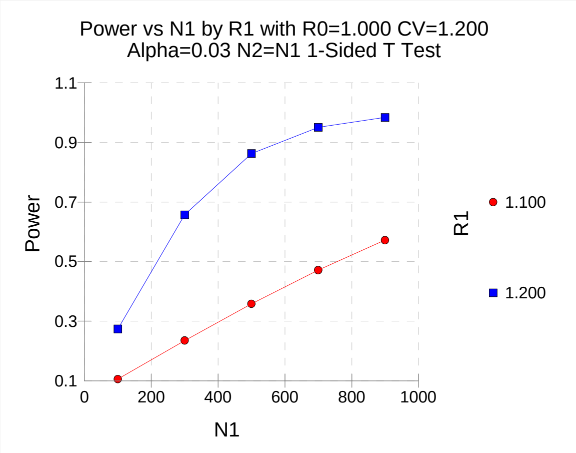
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sample T-Test Using Ratios
H0: R1=R0.  H1: R1>R0.

              Group         Mean         Mean             Coefficient
             Sample        Ratio        Ratio     Effect           of   Significance
              Sizes     Under H0     Under H1       Size    Variation          Level
Power       (N1/N2)         (R0)         (R1)       (ES)        (COV)        (Alpha)        Beta
0.1057      100/100        1.000        1.100     0.1009        1.200         0.0250      0.8943
0.2351      300/300        1.000        1.100     0.1009        1.200         0.0250      0.7649
0.3581      500/500        1.000        1.100     0.1009        1.200         0.0250      0.6419
0.4715      700/700        1.000        1.100     0.1009        1.200         0.0250      0.5285
0.5718      900/900        1.000        1.100     0.1009        1.200         0.0250      0.4282
0.2737      100/100        1.000        1.200     0.1930        1.200         0.0250      0.7263
0.6571      300/300        1.000        1.200     0.1930        1.200         0.0250      0.3429
0.8625      500/500        1.000        1.200     0.1930        1.200         0.0250      0.1375
0.9506      700/700        1.000        1.200     0.1930        1.200         0.0250      0.0494
0.9836      900/900        1.000        1.200     0.1930        1.200         0.0250      0.0164

Report Definitions
Power is the probability of rejecting a false null hypothesis. Power should be close to one.
N1 and N2 are the number of items sampled from each population.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.
R0 is the ratio of the means (Mean2/Mean1) under the null hypothesis, H0.
R1 is the ratio of the means (Mean2/Mean1) at which the power is calculated.
COV is the coefficient of variation on the original scale. The value of sigma is calculated from this.
ES is the effect size which is |Ln(R0)-Ln(R1)| / (sigma).

Summary Statements
A one-sided, two-sample t-test with group sample sizes of 100 and 100 achieves 11% power to
detect a ratio of 1.100 when the ratio under the null hypothesis is 1.000. The coefficent of
variation on the original scale is 1.200. The significance level (alpha) is 0.0250.
```

This report shows the power for the indicated scenarios.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Validation

We will validate this procedure by showing that it gives the identical results to the regular test on differences—a procedure that has been validated. We will use the same settings as those given in Example 1. Since the output for this example is shown above, only the output from the procedure that uses differences is shown below.

To run the power analysis of a *t-test* on differences, we need the values of Mean2 (which correspond to R1) and S1. The value of Mean1 will be zero.

$$\begin{aligned} S1 &= \sqrt{\ln\left(COV^2 + 1\right)} \\ &= \sqrt{\ln\left(1.2^2 + 1\right)} \\ &= 0.944456 \end{aligned}$$

$$\begin{aligned} Mean2 &= \ln\left(R1\right) \\ &= \ln\left(1.10\right) \\ &= 0.095310 \end{aligned} \qquad \begin{aligned} Mean2 &= \ln\left(R1\right) \\ &= \ln\left(1.20\right) \\ &= 0.182322 \end{aligned}$$

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means (Two-Sample T-Test) [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Specify using Differences (Two-Sample T-Test)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1a** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N1 (Sample Size Group 1)	**100 to 900 by 200**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**0.095310 0.182322**
S1 (Standard Deviation Group 1)	**0.944456**
S1 (Standard Deviation Group 2)	**S1**
Alternative Hypothesis	**Ha: Mean1<Mean2**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<Mean2
The standard deviations were assumed to be unknown and equal.

                              Allocation
Power        N1        N2        Ratio     Alpha      Beta     Mean1    Mean2       S1        S2
0.1057      100       100        1.000    0.0250    0.8943    0.0000   0.0953   0.9445    0.9445
0.2339      300       300        1.000    0.0250    0.7649    0.0000   0.0953   0.9445    0.9445
0.3581      500       500        1.000    0.0250    0.6419    0.0000   0.0953   0.9445    0.9445
0.4715      700       700        1.000    0.0250    0.5285    0.0000   0.0953   0.9445    0.9445
0.5718      900       900        1.000    0.0250    0.4282    0.0000   0.0953   0.9445    0.9445
0.2737      100       100        1.000    0.0250    0.7263    0.0000   0.1823   0.9445    0.9445
0.6556      300       300        1.000    0.0250    0.3429    0.0000   0.1823   0.9445    0.9445
0.8625      500       500        1.000    0.0250    0.1375    0.0000   0.1823   0.9445    0.9445
0.9506      700       700        1.000    0.0250    0.0494    0.0000   0.1823   0.9445    0.9445
0.9836      900       900        1.000    0.0250    0.0164    0.0000   0.1823   0.9445    0.9445
```

You can compare these power values with those shown above in Example 1 to validate the procedure. You will find that the power values are identical.

Chapter 450

Non-Inferiority & Superiority Tests for Two Means using Differences

Introduction

This procedure computes power and sample size for *non-inferiority* and *superiority* tests in two-sample designs in which the outcome is a continuous normal random variable. Measurements are made on individuals that have been randomly assigned to one of two groups. This is sometimes referred to as a *parallel-groups* design. This design is used in situations such as the comparison of the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs.

The two-sample t-test is commonly used with this situation. When the variances of the two groups are unequal, Welch's t-test may be used. When the data are not normally distributed, the Mann-Whitney (Wilcoxon signed-ranks) U test may be used.

The details of sample size calculation for the two-sample design are presented in the Two-Sample T-Test chapter and they will not be duplicated here. This chapter only discusses those changes necessary for non-inferiority and superiority tests. Sample size formulas for non-inferiority and superiority tests of two means are presented in Chow et al. (2003) pages 57-59.

The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests and their power and sample size could be calculated using the *Two-Sample T-Test* procedure. However, at the urging of our users, we have developed this module, which provides the input and output in formats that are convenient for these types of tests. This section will review the specifics of non-inferiority and superiority testing.

Remember that in the usual t-test setting, the null (H0) and alternative (H1) hypotheses for one-sided tests are defined as

$$\mathrm{H}_0{:}\,\mu_1-\mu_2\le D \text{ versus } \mathrm{H}_1{:}\,\mu_1-\mu_2> D$$

Rejecting this test implies that the mean difference is larger than the value *D*. This test is called an *upper-tailed test* because it is rejected in samples in which the difference between the sample means is larger than *D*.

Following is an example of a *lower-tailed test*.

$$\mathrm{H}_0{:}\,\mu_1-\mu_2\ge D \text{ versus } \mathrm{H}_1{:}\,\mu_1-\mu_2< D$$

Non-inferiority and *superiority* tests are special cases of the above directional tests. It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	**PASS Input/Output**	**Interpretation**
μ_1	Not used	*Mean* of population 1. Population 1 is assumed to consist of those who have received the new treatment.
μ_2	Not used	*Mean* of population 2. Population 2 is assumed to consist of those who have received the reference treatment.
ε	|E|	*Margin of equivalence.* This is a tolerance value that defines the magnitude of the amount that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value is shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	D	*True difference*. This is the value of $\mu_1-\mu_2$, the difference between the means. This is the value at which the power is calculated.

Note that the actual values of μ_1 and μ_2 are not needed. Only their difference is needed for power and sample size calculations.

Non-Inferiority and Superiority Tests

A *non-inferiority test* tests that the treatment mean is not worse than the reference mean by more than the equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

A *superiority test* tests that the treatment mean is better than the reference mean by more than the equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

Case 1: High Values Good, Non-Inferiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no less than a small amount below the reference mean. The value of δ is often set to zero. The following are equivalent sets of hypotheses.

$\mathrm{H}_0{:}\,\mu_1 \le \mu_2 - |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 > \mu_2 - |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_1 - \mu_2 \le -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 - \mu_2 > -|\varepsilon|$

$\mathrm{H}_0{:}\,\delta \le -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta > -|\varepsilon|$

Case 2: High Values Bad, Non-Inferiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no more than a small amount above the reference mean. The value of δ is often set to zero. The following are equivalent sets of hypotheses.

$\mathrm{H}_0{:}\,\mu_1 \ge \mu_2 + |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 < \mu_2 + |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_1 - \mu_2 \ge |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 - \mu_2 < |\varepsilon|$

$\mathrm{H}_0{:}\,\delta \ge |\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta < |\varepsilon|$

Case 3: High Values Good, Superiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is greater than the reference mean by at least the margin of equivalence. The value of δ must be greater than $|\varepsilon|$. The following are equivalent sets of hypotheses.

$\mathrm{H}_0{:}\,\mu_1 \le \mu_2 + |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 > \mu_2 + |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_1 - \mu_2 \le |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 - \mu_2 > |\varepsilon|$

$\mathrm{H}_0{:}\,\delta \le |\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta > |\varepsilon|$

Case 4: High Values Bad, Superiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is less than the reference mean by at least the margin of equivalence. The value of δ must be less than $-|\varepsilon|$. The following are equivalent sets of hypotheses.

$\mathrm{H}_0{:}\,\mu_1 \ge \mu_2 - |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 < \mu_2 - |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_1 - \mu_2 \ge -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_1 - \mu_2 < -|\varepsilon|$

$\mathrm{H}_0{:}\,\delta \ge -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta < -|\varepsilon|$

Example

A non-inferiority test example will set the stage for the discussion of the terminology that follows. Suppose that a test is to be conducted to determine if a new cancer treatment adversely affects mean bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 gm/cm with a standard deviation of 0.000300 gm/cm. Clinicians decide that if the treatment reduces AMBD by more than 5% (0.000115 gm/cm), it poses a significant health threat.

The hypothesis of interest is whether the mean AMBD in the treated group is more than 0.000115 below that of the reference group. The statistical test will be set up so that if the null hypothesis is rejected, the conclusion will be that the new treatment is non-inferior. The value 0.000115 gm/cm is called the *margin of equivalence* or the *margin of non-inferiority*.

Test Statistics

This section describes the test statistics that are available in this procedure.

Two-Sample T-Test

Under the null hypothesis, this test assumes that the two groups of data are simple random samples from a single population of normally-distributed values that all have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the test statistic for the case when higher response values are good is as follows.

$$t_{df} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - |\varepsilon|}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$\overline{X}_k = \frac{\sum_{i=1}^{N_k} X_{ki}}{N_k}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$df = N_1 + N_2 - 2$$

The null hypothesis is rejected if the computed p-value is less than a specified level (usually 0.05). Otherwise, no conclusion can be reached.

Welch's T-Test

Welch (1938) proposed the following test when the two variances are not assumed to be equal.

$$t_f^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - |\varepsilon|}{s^*_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s^*_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2}{N_1\left(N_1 - 1\right)}\right) + \left(\frac{\sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_2\left(N_2 - 1\right)}\right)}$$

$$f = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\left(N_1 - 1\right)} + \frac{s_2^4}{N_2^2\left(N_2 - 1\right)}}$$

$$s_1 = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2}{N_1 - 1}\right)}, \quad s_2 = \sqrt{\left(\frac{\sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_2 - 1}\right)}$$

Mann-Whitney U Test

This test is the nonparametric substitute for the equal-variance t-test. Two key assumptions are that the distributions are at least ordinal and that they are identical under H0. This means that ties (repeated values) are not acceptable. When ties are present, you can use approximations, but the theoretic results no longer hold.

The Mann-Whitney test statistic is defined as follows in Gibbons (1985).

$$z = \frac{W_1 - \frac{N_1(N_1 + N_2 + 1)}{2} + C}{s_W}$$

where

$$W_1 = \sum_{k=1}^{N_1} Rank\left(X_{1k}\right)$$

The ranks are determined after combining the two samples. The standard deviation is calculated as

$$s_W = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12} - \frac{N_1 N_2 \sum_{i=1} (t_i^3 - t_i)}{12(N_1 + N_2)(N_1 + N_2 - 1)}}$$

where t_i is the number of observations tied at value one, t_2 is the number of observations tied at some value two, and so forth.

The correction factor, *C*, is 0.5 if the rest of the numerator is negative or -0.5 otherwise. The value of *z* is then compared to the normal distribution.

Computing the Power

Standard Deviations Equal

When $\sigma_1 = \sigma_2 = \sigma$, the power of the *t* test is calculated as follows.

1. Find t_α such that $1 - T_{df}\left(t_\alpha\right) = \alpha$, where $T_{df}\left(t_\alpha\right)$ is the area under a central-*t* curve to the left of *x* and $df = N_1 + N_2 - 2$.
2. Calculate: $\sigma_{\bar{x}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$
3. Calculate the noncentrality parameter: $\lambda = \frac{|\varepsilon| - \delta}{\sigma_{\bar{x}}}$
4. Calculate: Power = $1 - T'_{df,\lambda}\left(t_\alpha\right)$, where $T'_{df,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ.

Standard Deviations Unequal

This case often recommends Welch's test. When $\sigma_1 \neq \sigma_2$, the power is calculated as follows.

1. Calculate: $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$.
2. Calculate: $f = \frac{\sigma_{\bar{x}}^4}{\frac{\sigma_1^4}{N_1^2(N_1+1)} + \frac{\sigma_2^4}{N_2^2(N_2+1)}} - 2$

   which is the adjusted degrees of freedom. Often, this is rounded to the next highest integer. Note that this is not the value of *f* used in the computation of the actual test. Instead, this is the expected value of *f*.
3. Find t_α such that $1 - T_f\left(t_\alpha\right) = \alpha$, where $T_f\left(t_\alpha\right)$ is the area to the left of *x* under a central-*t* curve with *f* degrees of freedom.
4. Calculate: $\lambda = \frac{|\varepsilon|}{\sigma_{\bar{x}}}$, the noncentrality parameter.
5. Calculate: Power = $1 - T'_{f,\lambda}\left(t_\alpha\right)$, where $T'_{f,\lambda}(x)$ is the area to the left of x under a noncentral-*t* curve with degrees of freedom *f* and noncentrality parameter λ.

Nonparametric Adjustment

When using the Mann-Whitney test rather than the *t* test, results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Mann-Whitney test may be made using the standard *t* test formulations with a simple adjustment to the sample sizes. The size of the adjustment depends on the actual distribution of the data. They give sample size adjustment factors for four distributions. These are 1 for uniform, 2/3 for double exponential, $9/\pi^2$ for logistic, and $\pi/3$ for normal distributions.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power and Beta* or *N1*.

Select *N1* when you want to determine the sample size needed to achieve a given power and alpha.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of inferiority when the null hypothesis should be rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of inferiority when in fact the mean is not non-inferior.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of group 1. Note that these values are ignored when you are solving for N1. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base N2 on the value of N1. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, N2 is calculated using the formula

  $$N2 = [R(N1)]$$

  where R is the Sample Allocation Ratio and the operator [Y] is the first integer greater than or equal to Y. For example, if you want N1 = N2, select *Use R* and set R = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when N2 is set to *Use R*.

When used, N2 is calculated from N1 using the formula: N2 = [R(N1)] where [Y] is the next integer greater than or equal to Y. Note that setting R = 1.0 forces N2 = N1.

Effect Size – Mean Difference

|E| (Equivalence Margin)

This is the magnitude of the *margin of equivalence*. It is the smallest difference between the mean and the reference mean that still results in the conclusion of non-inferiority (or superiority). Note that the sign of this value is assigned depending on the selections for Higher Is and Test Type.

D (True Difference)

This is the difference between the mean and the reference value at which the power is computed. For non-inferiority tests, this value is often set to zero, but it can be non-zero as long as the values are consistent with the alternative hypothesis, H1. For superiority tests, this value is non-zero. Again, it must be consistent with the alternative hypothesis, H1.

Effect Size – Standard Deviations

S1 and S2 (Standard Deviations)

These options specify the values of the standard deviations for each group. When the S2 is set to *S1*, only S1 needs to be specified. The value of S1 will be copied into S2.

When these values are not known, you must supply estimates of them. Press the *SD* button to display the Standard Deviation Estimator window. This procedure will help you find appropriate values for the standard deviation.

Test

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the mean is within the margin of equivalence of being no worse than the reference mean. Select *Superiority* when you want to test whether the mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are usually considered bad. However, if the response variable is income, higher values are generally considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Nonparametric Adjustment (Mann-Whitney Test)

This option makes appropriate sample size adjustments for the Mann-Whitney test. Results by Al-Sunduqchi and Guenther (1990) indicate that power calculations for the Mann-Whitney test may be made using the standard *t* test formulations with a simple adjustment to the sample size. The size of the adjustment depends upon the actual distribution of the data. They give sample size adjustment factors for four distributions. These are 1 for the uniform distribution, 2/3 for the double exponential distribution, $9/\pi^2$ for the logistic distribution, and $\pi/3$ for the normal distribution.

The options are as follows:

- **Ignore**

  Do not make a Mann-Whitney adjustment. This indicates that you want to analyze a *t* test, not the Wilcoxon test.

- **Uniform**

  Make the Mann-Whitney sample size adjustment assuming the uniform distribution. Since the factor is one, this option performs the same function as Ignore. It is included for completeness.

- **Double Exponential**

  Make the Mann-Whitney sample size adjustment assuming that the data actually follow the double exponential distribution.

- **Logistic**

  Make the Mann-Whitney sample size adjustment assuming that the data actually follow the logistic distribution.

- **Normal**

  Make the Mann-Whitney sample size adjustment assuming that the data actually follow the normal distribution.

Example 1 – Power Analysis

Suppose that a test is to be conducted to determine if a new cancer treatment adversely affects bone density. The adjusted mean bone density (AMBD) in the population of interest is 0.002300 gm/cm with a standard deviation of 0.000300 gm/cm. Clinicians decide that if the treatment reduces AMBD by more than 5% (0.000115 gm/cm), it poses a significant health threat. They also want to consider what would happen if the margin of equivalence is set to 2.5% (0.0000575 gm/cm).

Following accepted procedure, the analysis will be a non-inferiority test using the t-test at the 0.025 significance level. Power to be calculated assuming that the new treatment has no effect on AMBD. Several sample sizes between 10 and 800 will be analyzed. The researchers want to achieve a power of at least 90%. All numbers have been multiplied by 10000 to make the reports and plots easier to read.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N1 (Sample Size Group 1)	**10 50 100 200 300 500 600 800**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
S1 (Standard Deviation Group 1)	**3**
S2 (Standard Deviation Group 2)	**S1**
Test Type	**Non-Inferiority**
Higher is	**Good**
Nonparametric Adjustment	**Ignore**
Reports Tab	
Mean Decimals	**3**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

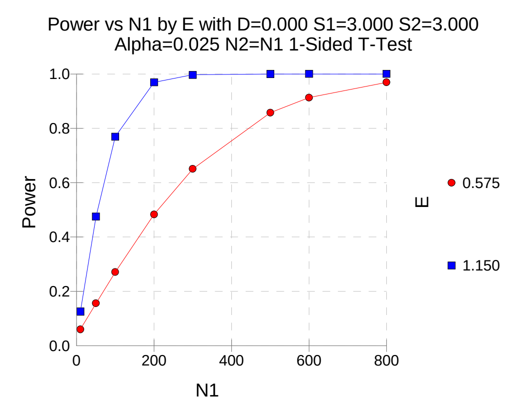
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

```
                        Equivalence     Actual  Significance                  Standard     Standard
                             Margin Difference         Level               Deviation1   Deviation2
Power            N1/N2          (E)        (D)       (Alpha)        Beta        (SD1)        (SD2)
0.06013          10/10       -0.575      0.000       0.02500     0.93987        3.000        3.000
0.15601          50/50       -0.575      0.000       0.02500     0.84399        3.000        3.000
0.27052        100/100       -0.575      0.000       0.02500     0.72948        3.000        3.000
0.48326        200/200       -0.575      0.000       0.02500     0.51674        3.000        3.000
0.65087        300/300       -0.575      0.000       0.02500     0.34913        3.000        3.000
0.85769        500/500       -0.575      0.000       0.02500     0.14231        3.000        3.000
0.91295        600/600       -0.575      0.000       0.02500     0.08705        3.000        3.000
0.96943        800/800       -0.575      0.000       0.02500     0.03057        3.000        3.000
0.12553          10/10       -1.150      0.000       0.02500     0.87447        3.000        3.000
0.47524          50/50       -1.150      0.000       0.02500     0.52476        3.000        3.000
0.76957        100/100       -1.150      0.000       0.02500     0.23043        3.000        3.000
0.96926        200/200       -1.150      0.000       0.02500     0.03074        3.000        3.000
0.99685        300/300       -1.150      0.000       0.02500     0.00315        3.000        3.000
0.99998        500/500       -1.150      0.000       0.02500     0.00002        3.000        3.000
1.00000        600/600       -1.150      0.000       0.02500     0.00000        3.000        3.000
1.00000        800/800       -1.150      0.000       0.02500     0.00000        3.000        3.000
```

Report Definitions
Group 1 is the treatment group. Group 2 is the reference or standard group.
Power is the probability of rejecting a false null hypothesis. Power should be close to one.
N1 and N2 are the sample sizes of group 1 and 2, respectively.
|E| is the magnitude of the margin of equivalence. It is the largest difference that is not of practical significance.
D is actual difference between the means. D = Mean1 - Mean2.
Alpha is the probability of a false-positive result.
Beta is the probability of a false-negative result.
SD1 and SD2 are the standard deviations of groups 1 and 2, respectively.

Summary Statements
Group sample sizes of 10 and 10 achieve 6% power to detect non-inferiority using a one-sided, two-sample t-test. The margin of equivalence is 0.575. The true difference between the means is assumed to be 0.000. The significance level (alpha) of the test is 0.02500. The data are drawn from populations with standard deviations of 3.000 and 3.000.

Chart Section



The above report shows that for |E| = 1.15, the sample size necessary to obtain 90% power is about 150 per group. However, if |E| = 0.575, the required sample size is about 600 per group.

Example 2 – Finding the Sample Size

Continuing with Example 1, the researchers want to know the exact sample size for each value of |E| to achieve 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.025**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
S1 (Standard Deviation Group 1)	**3**
S2 (Standard Deviation Group 2)	**S1**
Test Type	**Non-Inferiority**
Higher is	**Good**
Nonparametric Adjustment	**Ignore**
Reports Tab	
Mean Decimals	**3**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                      Equivalence    Actual  Significance              Standard     Standard
                           Margin  Difference        Level            Deviation1   Deviation2
Power           N1/N2         (E)        (D)       (Alpha)       Beta      (SD1)        (SD2)
0.90036       573/573      -0.575      0.000       0.02500    0.09964      3.000        3.000
0.90149       144/144      -1.150      0.000       0.02500    0.09851      3.000        3.000
```

This report shows the exact sample size requirement for each value of |E|.

Example 3 – Validation using Chow

Chow, Shao, Wang (2003) page 62 has an example of a sample size calculation for a non-inferiority trial. Their example obtains a sample size of 51 in each group when D = 0, |E| = 0.05, S = 0.1, Alpha = 0.05, and Beta = 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.80**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|E| (Equivalence Margin)	**0.05**
D (True Difference)	**0**
S1 (Standard Deviation Group 1)	**0.1**
S2 (Standard Deviation Group 2)	**S1**
Test Type	**Non-Inferiority**
Higher is	**Good**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                       Equivalence     Actual  Significance                Standard     Standard
                            Margin Difference         Level              Deviation1   Deviation2
Power            N1/N2         (E)        (D)       (Alpha)        Beta       (SD1)        (SD2)
0.80590          51/51      -0.050      0.000       0.05000     0.19410       0.100        0.100
```

PASS has also obtained a sample size of 51 per group.

Example 4 – Validation using Julious

Julious (2004) page 1950 gives an example of a sample size calculation for a parallel, non-inferiority design. His example obtains a sample size of 336 when D = 0, |E| = 10, S = 40, Alpha = 0.025, and Beta = 0.10.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.025**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|E| (Equivalence Margin)	**10**
D (True Difference)	**0**
S1 (Standard Deviation Group 1)	**40**
S2 (Standard Deviation Group 2)	**S1**
Test Type	**Non-Inferiority**
Higher is	**Good**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                         Equivalence     Actual  Significance                Standard      Standard
                              Margin Difference         Level             Deviation1    Deviation2
Power              N1/N2         (E)        (D)       (Alpha)        Beta      (SD1)         (SD2)
0.90045          337/337     -10.000      0.000       0.02500     0.09955     40.000        40.000
```

PASS obtained sample sizes of 337 in each group. The difference between 336 that Julious received and 337 that ***PASS*** calculated is likely caused by rounding.

Chapter 455

Non-Inferiority & Superiority Tests for Two Means using Ratios

Introduction

This procedure calculates power and sample size for *non-inferiority* and *superiority* t-tests from a parallel-groups design in which the logarithm of the outcome is a continuous normal random variable. This routine deals with the case in which the statistical hypotheses are expressed in terms of mean ratios instead of mean differences.

The details of testing the non-inferiority of two treatments using data from a two-group design are given in another chapter and they will not be repeated here. If the logarithm of the response can be assumed to follow a normal distribution, hypotheses about non-inferiority stated in terms of the ratio can be transformed into hypotheses about the difference. The details of this analysis are given in Julious (2004). They will only be summarized here.

Non-Inferiority Testing Using Ratios

It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	E	*Margin of equivalence*. This is a tolerance value that defines the maximum amount that is not of practical importance. This is the largest change in the mean ratio from the baseline value (usually one) that is still considered to be trivial.

Parameter	PASS Input/Output	Interpretation
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only the ratio of these values is needed for power and sample size calculations.

The null hypothesis of inferiority is

$$\mathrm{H}_0{:}\phi \le \phi_L \quad \text{where } \phi_L < 1.$$

and the alternative hypothesis of non-inferiority is

$$\mathrm{H}_1{:}\ \phi > \phi_L$$

Log-Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as scale-less percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of ratios.
2. Transform these into hypotheses about differences by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the null hypothesis are as follows.

$$\begin{aligned}&\phi_L \le \phi\\ &\Rightarrow \phi_L \le \left\{\frac{\mu_T}{\mu_R}\right\}\\ &\Rightarrow \ln(\phi_L) \le \left\{\ln(\mu_T) - \ln(\mu_R)\right\}\end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter can be used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable *X* is the logarithm of the original variable *Y*. That is, *X* = ln(*Y*) and *Y* = exp(*X*). Label the mean and variance of *X* as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of *Y* as μ_Y and σ_Y^2, respectively. If *X* is normally distributed, then *Y* is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be found to be

$$\begin{aligned} COV_Y &= \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y} \\ &= \sqrt{e^{\sigma_X^2} - 1} \end{aligned}$$

Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2 + 1 \right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

One final note: for parallel-group designs, σ_X^2 equals σ_d^2, the average variance used in the t-test of the logged data.

Thus, the hypotheses can be stated in the original (Y) scale and then the power can be analyzed in the transformed (X) scale.

Power Calculation

As is shown above, the hypotheses can be stated in the original (Y) scale using ratios or the logged (X) scale using differences. In either case, the power and sample size calculations are made using the formulas for testing the difference in two means. These formulas are presented in another chapter and are not duplicated here.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. In most situations, you will select either *Power and Beta* for a power analysis or *N1* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of inferiority when in fact the treatment mean is non-inferior.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when rejecting the null hypothesis of inferiority when in fact the treatment group is not inferior to the reference group.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 1 (the reference group). Note that these values are ignored when you are solving for N1. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 (the treatment group) or enter *Use R* to base N2 on the value of N1. You may enter a range of values such as *10 to 100 by 10*.

- **Use R**

  When *Use R* is entered here, N2 is calculated using the formula

  $$N2 = [R(N1)]$$

  where R is the Sample Allocation Ratio and the operator [Y] is the first integer greater than or equal to Y. For example, if you want N1 = N2, select *Use R* and set R = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when N2 is set to *Use R*.

When used, N2 is calculated from N1 using the formula: N2 = [R(N1)] where [Y] is the next integer greater than or equal to Y. Note that setting R = 1.0 forces N2 = N1.

Effect Size – Ratios

E (Equivalence Margin)

This is the magnitude of the relative *margin of equivalence*. It is the smallest change in the ratio of the two means that still results in the conclusion of non-inferiority (or superiority).

For example, suppose the non-inferiority boundary for the mean ratio is to be 0.80. This value is interpreted as follows: if the mean ratio (Treatment Mean / Reference Mean) is greater than 0.80, the treatment group is non-inferior to the reference group. In this example, the margin of equivalence would be 1.00 - 0.80 = 0.20.

This example assumes that higher values are better. If higher values are worse, an equivalence margin of 0.20 would be translated into a non-inferiority bound of 1.20. In this case, if the mean ratio is less than 1.20, the treatment group is non-inferior to the reference group.

Note that the sign of this value is ignored. Only the magnitude is used.

Recommended values:

0.20 is a common value for this parameter.

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Usually, the ratio will be assumed to be one. However, to be conservative, some authors recommend calculating the power using a ratio of 0.95 since this will require a larger sample size.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not logged) scale. This value must be determined from past experience or from a pilot study. See the discussion above for more details on the definition of the coefficient of variation.

Test

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the treatment mean is within the margin of equivalence of being no worse than the reference mean. Select *Superiority* when you want to test whether the treatment mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are usually considered bad. However, if the response variable is income, higher values are probably considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Example 1 – Finding Power

A company has developed a generic drug for treating rheumatism and wants to show that it is not inferior to the standard drug. Responses following either treatment are known to follow a log normal distribution. A parallel-group design will be used and the logged data will be analyzed with a two-sample t-test.

Researchers have decided to set the margin of equivalence at 0.20. Past experience leads the researchers to set the COV to 1.50. The significance level is 0.025. The power will be computed assuming that the true ratio is either 0.95 or 1.00. Sample sizes between 100 and 1000 will be included in the analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Non-Inferiority & Superiority Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N1 (Sample Size Group 1)	**100 to 1000 by 100**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
E (Equivalence Margin)	**0.20**
R1 (True Ratio)	**0.95 1.0**
COV (Coefficient of Variation)	**1.50**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Annotated Output

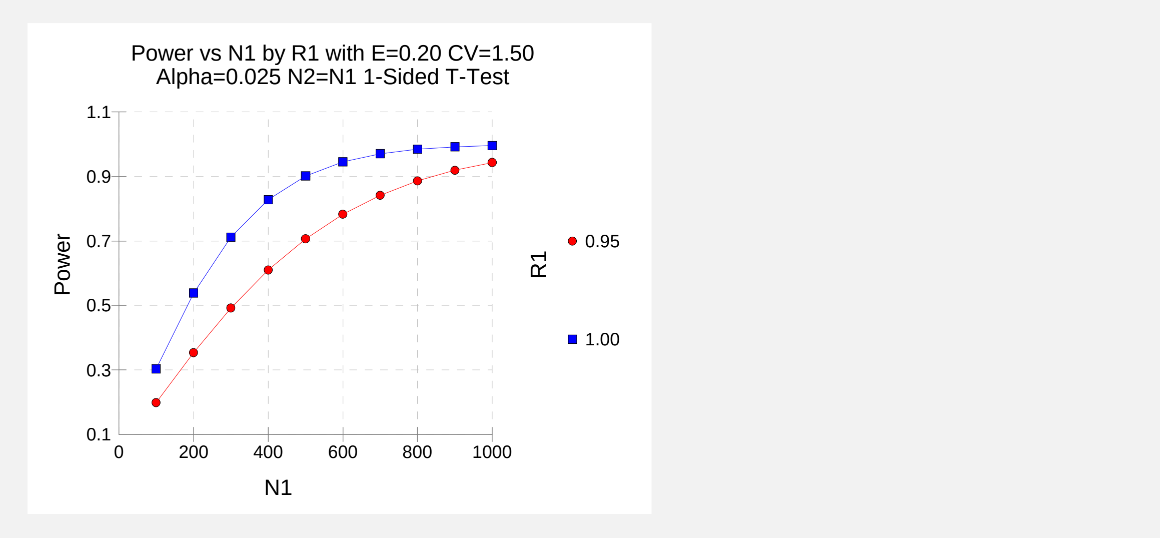
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Ratio Test (H0: R <= 1-E; H1: R > 1-E)

                          Equivalence  Equivalence     True  Significance  Coefficient
                               Margin        Bound    Ratio         Level  of Variation
Power           N1/N2             (E)         (RB)     (R1)       (Alpha)        (COV)       Beta
0.1987        100/100            0.20         0.80     0.95        0.0250         1.50     0.8013
0.3539        200/200            0.20         0.80     0.95        0.0250         1.50     0.6461
0.4918        300/300            0.20         0.80     0.95        0.0250         1.50     0.5082
0.6098        400/400            0.20         0.80     0.95        0.0250         1.50     0.3902
0.7064        500/500            0.20         0.80     0.95        0.0250         1.50     0.2936
0.7827        600/600            0.20         0.80     0.95        0.0250         1.50     0.2173
0.8416        700/700            0.20         0.80     0.95        0.0250         1.50     0.1584
0.8860        800/800            0.20         0.80     0.95        0.0250         1.50     0.1140
0.9189        900/900            0.20         0.80     0.95        0.0250         1.50     0.0811
0.9428      1000/1000            0.20         0.80     0.95        0.0250         1.50     0.0572
0.3038        100/100            0.20         0.80     1.00        0.0250         1.50     0.6962
0.5384        200/200            0.20         0.80     1.00        0.0250         1.50     0.4616
0.7113        300/300            0.20         0.80     1.00        0.0250         1.50     0.2887
0.8280        400/400            0.20         0.80     1.00        0.0250         1.50     0.1720
0.9013        500/500            0.20         0.80     1.00        0.0250         1.50     0.0987
0.9451        600/600            0.20         0.80     1.00        0.0250         1.50     0.0549
0.9702        700/700            0.20         0.80     1.00        0.0250         1.50     0.0298
0.9842        800/800            0.20         0.80     1.00        0.0250         1.50     0.0158
0.9918        900/900            0.20         0.80     1.00        0.0250         1.50     0.0082
0.9958      1000/1000            0.20         0.80     1.00        0.0250         1.50     0.0042

Report Definitions
H0 (null hypothesis) is that R <= 1-E, where R = Treatment Mean / Reference Mean.
H1 (alternative hypothesis) is that R > 1-E.
E is the magnitude of the relative margin of equivalence.
RB is equivalence bound for the ratio.
R1 is actual ratio between the treatment and reference means.
COV is the coefficient of variation on the original scale.
Power is the probability of rejecting H0 when it is false.
N1 is the number of subjects in the first (reference) group.
N2 is the number of subjects in the second (treatment) group.
Alpha is the probability of falsely rejecting H0.
Beta is the probability of not rejecting H0 when it is false.

Summary Statements
Group sample sizes of 100 in the first group and 100 in the second group achieve 20% power to
detect non-inferiority using a one-sided, two-sample t-test. The margin of equivalence is 0.20.
The true ratio of the means at which the power is evaluated is 0.95. The significance level
(alpha) of the test is 0.0250. The coefficients of variation of both groups are assumed to be
```

This report shows the power for the indicated scenarios.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Validation

We could not find a validation example for this procedure in the statistical literature. Therefore, we will show that this procedure gives the same results as the non-inferiority test on differences—a procedure that has been validated. We will use the same settings as those given in Example1. Since the output for this example is shown above, all that we need is the output from the procedure that uses differences.

To run the inferiority test on differences, we need the values of |E| and S1.

$$\begin{aligned} S1 &= \sqrt{\ln\left(COV^2+1\right)} \\ &= \sqrt{\ln\left(1.5^2+1\right)} \\ &= 1.085659 \end{aligned}$$

$$\begin{aligned} E' &= \ln\left(1-E\right) \\ &= \ln\left(0.8\right) \\ &= 0.223144 \end{aligned}$$

$$\begin{aligned} D &= \ln\left(R1\right) \\ &= \ln\left(0.95\right) \\ &= -0.051293 \end{aligned}$$

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1b** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N1 (Sample Size Group 1)	**100 to 1000 by 100**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|E| (Equivalence Margin)	**0.223144**
D (True Difference)	**-0.051293 0.0**
S1 (Standard Deviation Group 1)	**1.085659**
S2 (Standard Deviation Group 2)	**S1**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Test (H0: D <= -|E|; H1: D > -|E|)
Test Statistic: T-Test

                       Equivalence    Actual  Significance               Standard      Standard
                            Margin Difference        Level              Deviation1    Deviation2
Power           N1/N2          (E)       (D)       (Alpha)       Beta       (SD1)         (SD2)
0.1987        100/100       -0.223    -0.051        0.0250     0.8013       1.086         1.086
0.3539        200/200       -0.223    -0.051        0.0250     0.6461       1.086         1.086
0.4918        300/300       -0.223    -0.051        0.0250     0.5082       1.086         1.086
0.6098        400/400       -0.223    -0.051        0.0250     0.3902       1.086         1.086
0.7064        500/500       -0.223    -0.051        0.0250     0.2936       1.086         1.086
0.7828        600/600       -0.223    -0.051        0.0250     0.2172       1.086         1.086
0.8416        700/700       -0.223    -0.051        0.0250     0.1584       1.086         1.086
0.8860        800/800       -0.223    -0.051        0.0250     0.1140       1.086         1.086
0.9189        900/900       -0.223    -0.051        0.0250     0.0811       1.086         1.086
0.9428      1000/1000       -0.223    -0.051        0.0250     0.0572       1.086         1.086
0.3038        100/100       -0.223     0.000        0.0250     0.6962       1.086         1.086
0.5384        200/200       -0.223     0.000        0.0250     0.4616       1.086         1.086
0.7113        300/300       -0.223     0.000        0.0250     0.2887       1.086         1.086
0.8280        400/400       -0.223     0.000        0.0250     0.1720       1.086         1.086
0.9013        500/500       -0.223     0.000        0.0250     0.0987       1.086         1.086
0.9451        600/600       -0.223     0.000        0.0250     0.0549       1.086         1.086
0.9702        700/700       -0.223     0.000        0.0250     0.0298       1.086         1.086
0.9842        800/800       -0.223     0.000        0.0250     0.0158       1.086         1.086
0.9918        900/900       -0.223     0.000        0.0250     0.0082       1.086         1.086
0.9958      1000/1000       -0.223     0.000        0.0250     0.0042       1.086         1.086
```

You can compare these power values with those shown above in Example 1 to validate the procedure. You will find that the power values are identical.

Chapter 460

Equivalence Tests for Two Means using Differences

Introduction

This procedure allows you to study the power and sample size of equivalence tests of the means of two independent groups using the two-sample t-test. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. Only a brief introduction to the subject will be given here. For a comprehensive discussion, refer to Chow and Liu (1999).

Measurements are made on individuals that have been randomly assigned to one of two groups. This *parallel-groups* design may be analyzed by a TOST equivalence test to show that the means of the two groups do not differ by more than a small amount, called the margin of equivalence.

The definition of equivalence has been refined in recent years using the concepts of prescribability and switchability. *Prescribability* refers to ability of a physician to prescribe either of two drugs at the beginning of the treatment. However, once prescribed, no other drug can be substituted for it. *Switchability* refers to the ability of a patient to switch from one drug to another during treatment without adverse effects. Prescribability is associated with equivalence of location and variability. Switchability is associated with the concept of individual equivalence. This procedure analyzes average equivalence. Thus, it partially analyzes prescribability. It does not address equivalence of variability or switchability.

Parallel-Group Design

In a parallel-group design, subjects are assigned at random to either of two groups. Group 1 is the treatment group and group 2 is the reference group.

Outline of an Equivalence Test

PASS follows the *two one-sided tests* approach described by Schuirmann (1987) and Phillips (1990). Let $\mu_2 = \mu_T$ be the test group mean, $\mu_1 = \mu_R$ the reference group mean, and ε_L and ε_U the lower and upper bounds on $D = \mu_2 - \mu_1 = \mu_T - \mu_R$ that define the region of equivalence. The null hypothesis of non-equivalence is

$$H_0: D \le \varepsilon_L \quad or \quad H_0: D \ge \varepsilon_U$$

and the alternative hypothesis of equivalence is

$$H_1: \varepsilon_L < D < \varepsilon_U \,.$$

Two-Sample T-Test

This test assumes that the two groups of normally-distributed values have the same variance. The calculation of the two one-sided test statistics uses the following equations.

$$T_L = \frac{\left(\overline{X}_2 - \overline{X}_1\right) - \varepsilon_L}{s_{\overline{X}_1 - \overline{X}_2}} \text{ and } T_U = \frac{\left(\overline{X}_2 - \overline{X}_1\right) - \varepsilon_U}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$\overline{X}_k = \frac{\sum_{i=1}^{N_k} X_{ki}}{N_k}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$df = N_1 + N_2 - 2$$

The null hypothesis is rejected if T_L and $-T_U$ are greater than or equal to $t_{1-\alpha, N_1 + N_2 - 2}$.

The power of this test is given by

$$\Pr(T_L \ge t_{1-\alpha,N_1+N_2-2} \;\; and \;\; T_U \le -t_{1-\alpha,N_1+N_2-2} | \mu_T, \mu_R, \sigma^2)\, 1$$

where T_L and T_U are distributed as the bivariate, noncentral *t* distribution with noncentrality parameters Δ_L and Δ_U given by

$$\Delta_L = \frac{D - \varepsilon_L}{\sigma\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$\Delta_U = \frac{D - \varepsilon_U}{\sigma\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N1* for sample size determination.

Select *N1* when you want to calculate the sample size needed to achieve a given power and alpha level.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of nonequivalent means when in fact the means are equivalent.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of non-equivalent means when in fact the means are nonequivalent.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Reference Group)

Specify the number of subjects in the reference group. The total number of subjects in the experiment is equal to *N1* + *N2*.

You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Treatment Group)

Specify one or more values for the number of subjects in the treatment group. Alternatively, enter *Use R* to base *N2* on the value of *N1*. You may also enter a range of values such as *10 to 100 by 10*.

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and [*Y*] means take the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: $N2= [R(N1)]$ where [*Y*] means take the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N1* = *N2*.

Effect Size – Equivalence Limits

|EU| Upper Equivalence Limit

This value gives upper limit on equivalence. Differences outside EL and EU are not considered equivalent. Differences between them are considered equivalent.

Note that EL<0 and EU>0. Also, you must have EL<D<EU.

-|EL| Lower Equivalence Limit

This value gives lower limit on equivalence. Differences outside EL and EU are not considered equivalent. Differences between them are.

If you want symmetric limits, enter -UPPER LIMIT for EL to force EL = -|EU|.

Note that EL<0 and EU>0. Also, you must have EL<D<EU. Finally, the scale of these numbers must match the scale of S.

Effect Size – True Mean Difference

D (True Difference)

This is the true difference between the two means at which the power is to be computed. Often this value is set to zero, but it can be non-zero as long as it is between the equivalence limits EL and EU.

Effect Size – Standard Deviation

S (Standard Deviation)

Specify the within-group standard deviation, σ . The standard deviation is assumed to be the same for both groups.

Example 1 – Parallel-Group Design

A parallel-group is to be used to compare influence of two drugs on diastolic blood pressure. The diastolic blood pressure is known to be close to 96 mmHg with the reference drug and is thought to be 92 mmHg with the experimental drug. Based on similar studies, the within-group standard deviation is set to 18mmHg. Following FDA guidelines, the researchers want to show that the diastolic blood pressure with the experimental drug is within 20% of the diastolic blood pressure with the reference drug. Note that 20% of 96 is 19.2. They decide to calculate the power for a range of sample sizes between 3 and 60. The significance level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**3 5 8 10 15 20 30 40 50 60**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|EU| Upper Equivalence Limit	**19.2**
-|EL| Lower Equivalence Limit	**-Upper Limit**
D (True Difference)	**-4**
S (Standard Deviation)	**18**

Annotated Output

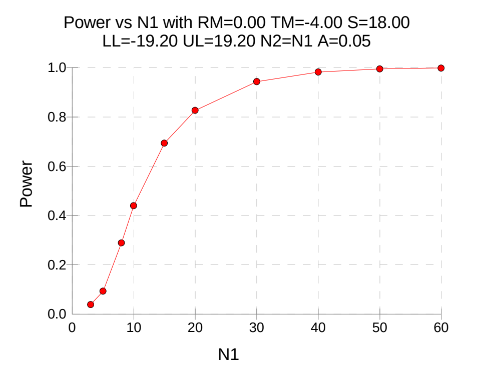
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Parallel-Group Design

         Reference  Treatment
             Group      Group
            Sample     Sample     Lower     Upper
              Size       Size    Equiv.    Equiv.       True   Standard
Power         (N1)       (N2)     Limit     Limit Difference  Deviation      Alpha       Beta
0.0386           3          3    -19.20     19.20      -4.00      18.00     0.0500     0.9614
0.0928           5          5    -19.20     19.20      -4.00      18.00     0.0500     0.9072
0.2887           8          8    -19.20     19.20      -4.00      18.00     0.0500     0.7113
0.4391          10         10    -19.20     19.20      -4.00      18.00     0.0500     0.5609
0.6934          15         15    -19.20     19.20      -4.00      18.00     0.0500     0.3066
0.8266          20         20    -19.20     19.20      -4.00      18.00     0.0500     0.1734
0.9433          30         30    -19.20     19.20      -4.00      18.00     0.0500     0.0567
0.9820          40         40    -19.20     19.20      -4.00      18.00     0.0500     0.0180
0.9946          50         50    -19.20     19.20      -4.00      18.00     0.0500     0.0054
0.9984          60         60    -19.20     19.20      -4.00      18.00     0.0500     0.0016

Report Definitions
Power is the probability of rejecting non-equivalence when they are equivalent.
N1 is the number of subjects in the reference group.
N2 is the number of subjects in the treatment group.
The Upper & Lower Limits are the maximum allowable differences that result in equivalence.
True Difference is the anticipated actual difference between the means.
The Standard Deviation is the average S.D. within the two groups.
Alpha is the probability of rejecting non-equivalence when they are non-equivalent.
Beta is the probability of accepting non-equivalence when they are equivalent.

Summary Statements
An equivalence test of means using two one-sided tests on data from a parallel-group design
with sample sizes of 3 in the reference group and 3 in the treatment group achieves 4% power at
a 5% significance level when the true difference between the means is -4.00, the standard
deviation is 18.00, and the equivalence limits are -19.20 and 19.20.
```

This report shows the power for the indicated parameter configurations. Note that when the parameters are specified as percentages, they are displayed in the output with percent signs. Note that the desired 80% power occurs for a per group sample size between 15 and 20.

Plot Section



This plot shows the power versus the sample size.

Example 2 – Parallel-Group Validation using Machin

Machin *et al.* (1997) page 107 present an example of determining the sample size for a parallel-group design in which the reference mean is 96, the treatment mean is 94, the standard deviation is 8, the limits are plus or minus 5, the power is 80%, and the significance level is 0.05. They calculate the sample size to be 88. It is important to note that Machin *et al.* use an approximation, so their results should not be expected to exactly match the results obtained using ***PASS***.

We will now set up this example in ***PASS***.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.80**
Alpha	**0.05**
N1 (Sample Size Group 1)	**Ignored**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
|EU| Upper Equivalence Limit	**5**
-|EL| Lower Equivalence Limit	**-Upper Limit**
D (True Difference)	**-2**
S (Standard Deviation)	**8**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
          Reference  Treatment
              Group      Group
             Sample     Sample     Lower     Upper
               Size       Size    Equiv.    Equiv.       True   Standard
Power          (N1)       (N2)     Limit     Limit Difference  Deviation      Alpha        Beta
0.8015           89         89     -5.00      5.00      -2.00       8.00     0.0500      0.1985
```

Note that ***PASS*** has obtained a sample size of 89 which is very close to the approximate value of 88 that Machin calculated.

Chapter 465

Equivalence Tests for Two Means (Simulation)

Introduction

This procedure allows you to study the power and sample size of an equivalence test comparing two means from independent groups. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. The t-test is commonly used in this situation, but other tests have been developed for use when the t-test assumptions are not met. These additional tests include the Mann-Whitney U test, Welch's unequal variance test, and trimmed versions of the t-test and the Welch test.

Measurements are made on individuals that have been randomly assigned to, or randomly chosen from, one of two groups. This *parallel-groups* design may be analyzed by a TOST equivalence test to show that the means of the two groups do not differ by more than a small amount, called the margin of equivalence.

The two-sample t-test is commonly used in this situation. When the variances of the two groups are unequal, Welch's t-test is often used. When the data are not normally distributed, the Mann-Whitney (Wilcoxon signed-ranks) U test and, less frequently, the trimmed t-test may be used.

The details of the power analysis of equivalence test using analytic techniques are presented in another ***PASS*** chapter and they won't be duplicated here. This chapter will only consider power analysis using computer simulation.

Technical Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are as follows.

1. Specify how the test is carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.

2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's power.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's significance level.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in ***PASS*** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draw the random numbers from this pool. This second method can cut the running time of the simulation by 70%.

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Simulating Data for an Equivalence Test

Simulating equivalence data is more complex than simulating data for a regular two-sided test. An equivalence test essentially reverses the roles of the null and alternative hypothesis. The null hypothesis becomes

$$H0\colon (\mu_1 - \mu_2) \le -D \text{ or } (\mu_1 - \mu_2) \ge D$$

where *D* is the margin of equivalence. Thus the null hypothesis is made up of two simple hypotheses:

$$H0_1\colon (\mu_1 - \mu_2) \le -D$$

$$H0_2\colon (\mu_1 - \mu_2) \ge D$$

The additional complexity comes in deciding which of the two null hypotheses are used to simulate data for the null hypothesis situation. The choice becomes more problematic when asymmetric equivalence limits are chosen. In this case, you may want to try simulating using each simple null hypothesis in turn.

To generate data for the null hypotheses, generate data for each group. The difference in the means of these two groups will become one of the equivalence limits. The other equivalence limit will be determined by symmetry and will always have a sign that is the opposite of the first equivalence limit.

Test Statistics

This section describes the test statistics that are available in this procedure. Note that these test statistics are computed on the differences. Thus, when the equation refers to an X value, this X value is assumed to be a difference between two individual variates.

Two-Sample T-Test

The t-test assumes that the data are simple random samples from populations of normally-distributed values that have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the t statistic is as follows.

$$t_{df} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$\overline{X}_k = \frac{\sum_{i=1}^{N_k} X_{ki}}{N_k}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$df = N_1 + N_2 - 2$$

The significance of the test statistic is determined by computing a p-value which is based on the t distribution with appropriate degrees of freedom. If this p-value is less than a specified level (often 0.05), the null hypothesis is rejected. Otherwise, no conclusion can be reached.

Welch's T-Test

Welch (1938) proposed the following test for use when the two variances are not assumed to be equal.

$$t_f^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{s^*_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s^*_{\overline{X}_1-\overline{X}_2}=\sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i}-\overline{X}_1\right)^2}{N_1\left(N_1-1\right)}\right)+\left(\frac{\sum_{i=1}^{N_2}\left(X_{2i}-\overline{X}_2\right)^2}{N_2\left(N_2-1\right)}\right)}$$

$$f=\frac{\left(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\left(N_1-1\right)}+\frac{s_2^4}{N_2^2\left(N_2-1\right)}}$$

$$s_1=\sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i}-\overline{X}_1\right)^2}{N_1-1}\right)}\,,\quad s_2=\sqrt{\left(\frac{\sum_{i=1}^{N_2}\left(X_{2i}-\overline{X}_2\right)^2}{N_2-1}\right)}$$

Trimmed T-Test assuming Equal Variances

The notion of trimming off a small proportion of possibly outlying observations and using the remaining data to form a t-test was first proposed for one sample by Tukey and McLaughlin (1963). Tukey and Dixon (1968) consider a slight modification of this test, called *Winsorization*, which replaces the trimmed data with the nearest remaining value. The two-sample trimmed t-test was proposed by Yuen and Dixon (1973).

Assume that the data values have been sorted from lowest to highest. The *trimmed mean* is defined as

$$\overline{X}_{tg}=\frac{\sum_{k=g+1}^{N-g}X_k}{h}$$

where $h = N - 2g$ and $g = [N(G/100)]$. Here we use $[Z]$ to mean the largest integer smaller than Z with the modification that if G is non-zero, the value of $[N(G/100)]$ is at least one. G is the percent trimming and should usually be less than 25%, often between 5% and 10%. Thus, the g smallest and g largest observation are omitted in the calculation.

To calculate the modified t-test, calculate the *Winsorized mean* and the *Winsorized* sum of squared deviations as follows.

$$\overline{X}_{wg}=\frac{g\left(X_{g+1}+X_{N-g}\right)+\sum_{k=g+1}^{N-g}X_k}{N}$$

$$SSD_{wg}=\frac{g\left(X_{g+1}-\overline{X}_{wg}\right)^2+g\left(X_{N-g}-\overline{X}_{wg}\right)^2+\sum_{k=g+1}^{N-g}\left(X_k-\overline{X}_{wg}\right)^2}{N}$$

Using the above definitions, the two-sample trimmed t-test is given by

$$T_{tg} = \frac{\left(\overline{X}_{1tg} - \overline{X}_{2tg}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{SSD_{1wg} + SSD_{2wg}}{h_1 + h_2 - 2}\left(\frac{1}{h_1} + \frac{1}{h_2}\right)}}$$

The distribution of this *t* statistic is approximately that of a *t* distribution with degrees of freedom equal to $h_1 + h_2 - 2$. This approximation is often reasonably accurate if both sample sizes are greater than 6.

Trimmed T-Test assuming Unequal Variances

Yuen (1974) combines trimming (see above) with Welch's (1938) test. The resulting trimmed Welch test is resistant to outliers and seems to alleviate some of the problems that occur because of skewness in the underlying distributions. Extending the results from above, the trimmed version of Welch's t-test is given by

$$T_{tg}^{*} = \frac{\left(\overline{X}_{1tg} - \overline{X}_{2tg}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)} + \frac{SSD_{2wg}}{h_2\left(h_2 - 1\right)}}}$$

with degrees of freedom *f* given by

$$\frac{1}{f} = \frac{c^2}{h_1 - 1} + \frac{1 - c^2}{h_2 - 1}$$

where

$$c = \frac{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)}}{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)} + \frac{SSD_{2wg}}{h_2\left(h_2 - 1\right)}}$$

Mann-Whitney U Test

This test is the nonparametric substitute for the equal-variance t-test. Two key assumptions for this test are that the distributions are at least ordinal and that they are identical under H0. This means that ties (repeated values) are not acceptable. When ties are present, an approximation can be used, but the theoretic results no longer hold.

The Mann-Whitney test statistic is defined as follows in Gibbons (1985).

$$z = \frac{W_1 - \frac{N_1(N_1 + N_2 + 1)}{2} + C}{s_W}$$

where

$$W_1 = \sum_{k=1}^{N_1} Rank\left(X_{1k}\right)$$

The ranks are determined after combining the two samples. The standard deviation is calculated as

$$s_W = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12} - \frac{N_1 N_2 \sum_{i=1} (t_i^3 - t_i)}{12(N_1 + N_2)(N_1 + N_2 - 1)}}$$

where t_i is the number of observations tied at value one, t_2 is the number of observations tied at some value two, and so forth.

The correction factor, *C*, is 0.5 if the rest of the numerator of *z* is negative or -0.5 otherwise. The value of *z* is then compared to the standard normal distribution.

Standard Deviations

Care must be used when either the null or alternative distribution is not normal. In these cases, the standard deviation is usually not specified directly. For example, you might use a gamma distribution with a shape parameter of 1.5 and a mean of 4 as the null distribution and a gamma distribution with the same shape parameter and a mean of 5 as the alternative distribution. This allows you to compare the two means. However, although the shape parameters are constant, the standard deviations, which are based on both the shape parameter and the mean, are not. Thus the distributions not only have different means, but different standard deviations!

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data and Options tabs. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated using the values of the other parameters. Under most conditions, you would select either *Power* or *N1*.

Select *Power* when you want to estimate the power for a specific scenario.

Select *N1* when you want to determine the sample size needed to achieve a given power and alpha level. This option is computationally intensive and may take a long time to complete.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of nonequivalent means when in fact the means are equivalent.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of non-equivalent means when in fact the means are nonequivalent.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of group 1. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

Use R

When *Use R* is entered here, *N2* is calculated using the formula

$$N2 = [R(N1)]$$

where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: $N2 = [R(N1)]$ where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Test

Test Type

Specify which test statistic is to be used in the simulation. Although the t-test is the most commonly used test statistic, it is based on assumptions that may not be viable in many situations. For your data, you may find that one of the other tests are more accurate (actual alpha = target alpha) and more precise (better power).

Equivalence Limit

Equivalence limits are defined as the positive and negative limits around zero that define a zone of equivalence. This zone of equivalence is a set of difference values that define a region in which the two means are ‘close enough’ so that they are considered to be the same for practical purposes.

Rather than define these limits explicitly, they are set implicitly. This is done as follows. One limit is found by subtracting the Group 2 Dist’n|H0 mean from the Group 1 Dist’n|H0 mean. If the limits are symmetric, the other limit is this difference times -1. To obtain symmetric limits, enter ‘Symmetric’ here.

If asymmetric limits are desired, a numerical value is specified here. It is given the sign (+ or -) that is opposite the difference of the means discussed above.

For example, if the mean of group 1 under H0 is 5, the mean of group 2 under H0 is 4, and *Symmetric* is entered here, the equivalence limits will be 5 - 4 = 1 and -1. However, if the value *1.25* is entered here, the equivalence limits are 1 and -1.25.

If you do not have a specific value in mind for the equivalence limit, a common value for an equivalence limit is 20% or 25% of the group 1 (reference) mean.

Simulations

Simulations

This option specifies the number of iterations, *M*, used in the simulation. The larger the number of iterations, the longer the running time, and, the more accurate the results.

The precision of the simulated power estimates are calculated from the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the ‘Precision’ amount given in the table.

Simulation Size M	Precision when Power = 0.50	Precision when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

Group 1 (and 2) Distribution | H0

These options specify the distributions of the two groups under the null hypothesis, H0. The difference between the means of these two distributions is the value of one of the equivalence limits.

Group 1 is often called the reference (or standard) distribution. Group 2 is often called the treatment distribution. These options specify these two distributions under the null hypothesis, H0. The difference between the means of these two distributions is, by definition, one of the equivalence limits. Thus, you set the equivalence limit by specifying the two means.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The characters *M0* and *M1* are to be used for the means of the distributions of groups 1 and 2 under H0, respectively. An equivalence limit is then *M0 - M1*, which must be non-zero.

For example, suppose you entered *N(M0 S)* for group 1 and *N(M1 S)* for group 2. Also, you set *M0* equal to 5 and M1 equal to 4. The upper (positive) equivalence limit would be 5 – 4 = 1.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean is entered first.

```
Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)
```

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean of a Specified Distribution

The distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5,

6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Group 1 (and 2) Distribution|H1

These options specify the distributions of the two groups under the alternative hypothesis, H1. The difference between the means of these two distributions is the difference that is assumed to be the true value of the difference.

Usually, the mean difference is specified by entering *M0* for the mean parameter in the distribution expression for group 1 and *M1* for the mean parameter in the distribution expression for group 2. The mean difference under H1 then becomes the value of *M0– M0 = 0*. If you want a non-zero value, you specify it by specifying unequal values for the two distribution means. For example, you could enter *A* for the mean of group 2. The mean difference will then be *M0 – A*.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean of group 2 under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean, *M1*, is entered first.

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for *M0* in the distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for *M1* in the distribution specifications given above. Although it can be used wherever you want, *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B)

Enter the numeric value(s) of parameter listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values using the syntax ‘0 2 3’ or ‘0 to 3 by 1.’

You can also change the letter than is used as the name of this parameter.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size, N1, is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Random Numbers

Random Number Pool Size

This is the size of the pool of values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Trimmed T-Test

Percent Trimmed at Each End

Specify the percent of each end of the sorted data that is to be trimmed (constant *G* above) when using the trimmed means procedures. This percentage is applied to the sample size to determine how many of the lowest and highest data values are to be trimmed by the procedure. For example, if the sample size (N1) is 27 and you specify 10 here, then [27*10/100] = 2 observations will be trimmed at the bottom and the top. For any percentage, at least one observation is trimmed from each end of the sorted dataset.

The range of possible values is 0 to 25.

Example 1 – Power at Various Sample Sizes

Researchers are planning an experiment to determine if the response to a new drug is equivalent to the response to the standard drug. The average response level to the standard drug is known to be 63 with a standard deviation of 5. The researchers decide that if the average response level to the new drug is between 60 and 66, they will consider it to be equivalent to the standard drug.

The researchers decide to use a parallel-group design. The response level for the standard drug will be measured for each subject. They will analyze the data using an equivalence test based on the t-test with an alpha level of 0.05. They want to compare the power at sample sizes of 10, 30, 50, and 70. They assume that the data are normally distributed and that the true difference between the mean response of the two drugs is zero. Since this is an exploratory analysis, the number of simulation iterations is set to 2000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**10 30 50 70**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Group 1 Distribution | H0	**N(M0 S)**
Group 2 Distribution | H0	**N(M1 S)**
Group 1 Distribution | H1	**N(M0 S)**
Group 2 Distribution | H1	**N(M0 S)**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

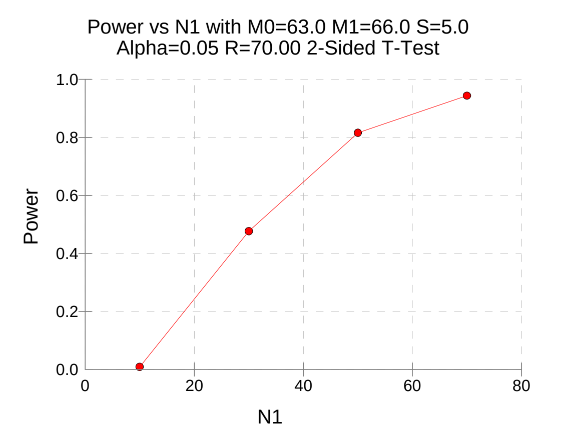
Numeric Results and Plots

```
Numeric Results for Testing Mean Equivalence.    Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S) & Normal(M1 S)
H1 Dist's: Normal(M0 S) & Normal(M0 S)
Test Statistic: T-Test
                          Lower   Upper
                   H1    Equiv.  Equiv.  Target  Actual
Power    N1/N2   Diff1    Limit   Limit   Alpha   Alpha    Beta      M0      M1      S
0.009    10/10     0.0     -3.0     3.0   0.050   0.005   0.991    63.0    66.0    5.0
(0.004)  [0.005  0.013]                          (0.003)  [0.002  0.008]

0.477    30/30     0.0     -3.0     3.0   0.050   0.053   0.524    63.0    66.0    5.0
(0.022)  [0.455  0.498]                          (0.010)  [0.043  0.062]

0.816    50/50     0.0     -3.0     3.0   0.050   0.061   0.184    63.0    66.0    5.0
(0.017)  [0.799  0.833]                          (0.010)  [0.050  0.071]

0.944    70/70     0.0     -3.0     3.0   0.050   0.050   0.056    63.0    66.0    5.0
(0.010)  [0.934  0.954]                          (0.010)  [0.040  0.060]

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 21.61 seconds.
```

Summary Statements

Group sample sizes of 10 and 10 achieve 1% power to detect equivalence when the margin of equivalence is from -3.0 to 3.0 and the actual mean difference is 0.0. The significance level (alpha) is 0.050 using two one-sided T-Tests. These results are based on 2000 Monte Carlo samples from the null distributions: Normal(M0 S) and Normal(M1 S), and the alternative distributions: Normal(M0 S) and Normal(M0 S).

Chart Section



This report shows the estimated power for each scenario. The first row shows the parameter settings and the estimated power and significance level (Actual Alpha). The second row shows two 95% confidence intervals in brackets: the first for the power and the second for the significance level. Half the width of each confidence interval is given in parentheses as a fundamental measure of the accuracy of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Example 2 – Finding the Sample Size

Continuing with Example1, the researchers want to determine how large a sample is needed to obtain a power of 0.90.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Group 1 Distribution | H0	**N(M0 S)**
Group 2 Distribution | H0	**N(M1 S)**
Group 1 Distribution | H1	**N(M0 S)**
Group 2 Distribution | H1	**N(M0 S)**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Mean Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S) & Normal(M1 S)
H1 Dist's: Normal(M0 S) & Normal(M0 S)
Test Statistic: T-Test
                            Lower    Upper
                     H1     Equiv.   Equiv.   Target   Actual
Power    N1/N2      Diff1   Limit    Limit    Alpha    Alpha      Beta       M0        M1         S
0.911    61/61       0.0     -3.0      3.0    0.050    0.044     0.089     63.0      66.0       5.0
(0.012)  [0.899    0.923]                              (0.009)   [0.035    0.053]
```

The required sample size is 61 per group.

Example 3 – Comparative Results when the Data Contain Outliers

Continuing Example1, this example will investigate the impact of outliers on the characteristics of the various test statistics. The two-sample t-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracy when the data contains outliers. This example will investigate the impact of outliers on the power and precision of the five test statistics available in ***PASS***.

A mixture of two normal distributions will be used to randomly generate outliers. The mixture will draw 95% of the data from a standard distribution. The other 5% of the data will come from a normal distribution with the same mean but with a standard deviation that is one, five, and ten times larger than that of the standard.

Setup

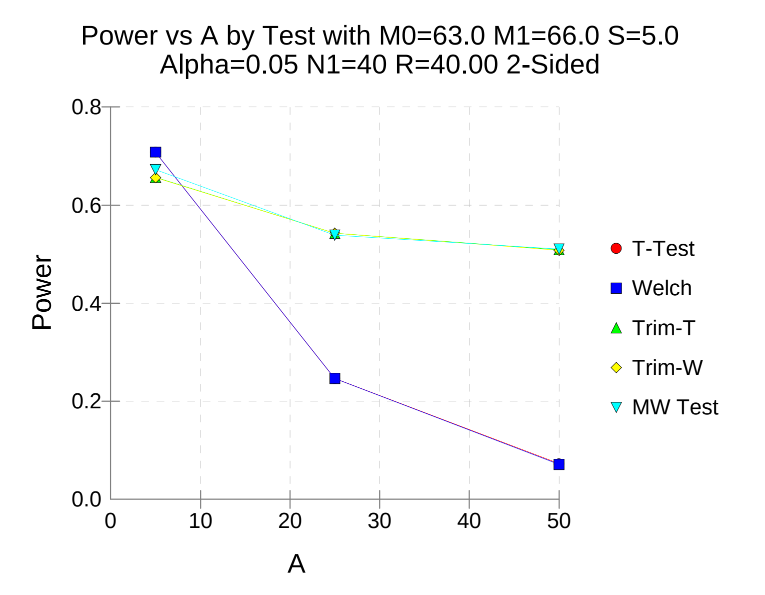
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**40**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Group 1 Distribution | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Distribution | H0	**N(M1 S)[95];N(M1 A)[5]**
Group 1 Distribution | H1	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Distribution | H1	**N(M0 S)[95];N(M0 A)[5]**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**
A	**5 25 50**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

```
Power Comparison for Testing Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M1 S)[95];Normal(M1 A)[5]
H1 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M0 S)[95];Normal(M0 A)[5]

          H1    Lower   Upper                         Trim.   Trim.   Mann
        Diff   Equiv.  Equiv.  Target  T-Test  Welch T-Test   Welch  Whit'y
N1/N2 (Diff1)   Limit   Limit   Alpha   Power  Power  Power   Power   Power     M0     M1     S      A
40/40    0.0     -3.0     3.0   0.050   0.708  0.708  0.657   0.656   0.672   63.0   66.0   5.0    5.0
40/40    0.0     -3.0     3.0   0.050   0.247  0.247  0.543   0.543   0.539   63.0   66.0   5.0   25.0
40/40    0.0     -3.0     3.0   0.050   0.073  0.072  0.509   0.508   0.510   63.0   66.0   5.0   50.0

Alpha Comparison for Testing Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|

          H1    Lower   Upper                         Trim.   Trim.   Mann
        Diff   Equiv.  Equiv.  Target  T-Test  Welch T-Test   Welch  Whit'y
N1/N2 (Diff1)   Limit   Limit   Alpha   Alpha  Alpha  Alpha   Alpha   Alpha     M0     M1     S      A
40/40    0.0     -3.0     3.0   0.050   0.050  0.050  0.058   0.058   0.056   63.0   66.0   5.0    5.0
40/40    0.0     -3.0     3.0   0.050   0.030  0.030  0.041   0.041   0.044   63.0   66.0   5.0   25.0
40/40    0.0     -3.0     3.0   0.050   0.008  0.008  0.042   0.042   0.044   63.0   66.0   5.0   50.0
Pool Size: 10000. Simulations: 2000. Run Time: 2.90 minutes. Percent Trimmed: 10.
```



When A = 5, there are no outliers and the power of the nonparametric test and the trimmed tests are a little less than that of the t-test. When A = 25, the distortion of the t-test caused by the outliers becomes apparent. In this case, the powers of the standard t-test and Welch's t-test are 0.247, but the powers of the nonparametric Mann-Whitney test and the trimmed tests are about 0.54. When A = 50, the standard t-test only achieves a power of 0.073, but the trimmed and nonparametric tests achieve powers of about 0.51!

Looking at the second table, we see that the true significance level of the t-test is distorted by the outliers, while the significance levels of the other tests remain close to the target value.

Example 4 – Selecting a Test Statistic when the Data Are Skewed

Continuing Example3, this example will investigate the impact of skewness in the underlying distribution on the characteristics of the various test statistics.

Tukey’s lambda distribution will be used because it allows the amount of skewness to be gradually increased.

Setup

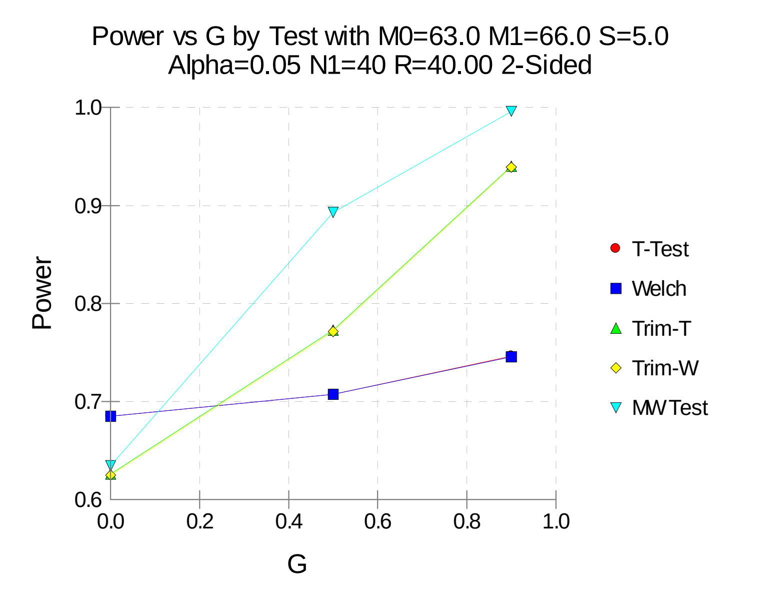
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**40**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Group 1 Distribution | H0	**L(M0 S G 0)**
Group 2 Distribution | H0	**L(M1 S G 0)**
Group 1 Distribution | H1	**L(M0 S G 0)**
Group 2 Distribution | H1	**L(M0 S G 0)**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**
G	**0 0.5 0.9**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

```
Power Comparison for Testing Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M1 S)[95];Normal(M1 A)[5]
H1 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M0 S)[95];Normal(M0 A)[5]

             H1    Lower   Upper                         Trim.   Trim.   Mann
           Diff   Equiv.  Equiv.  Target  T-Test  Welch T-Test  Welch  Whit'y
N1/N2  (Diff1)    Limit   Limit   Alpha   Power   Power  Power  Power   Power     M0     M1     S     G
40/40      0.0     -3.0     3.0   0.050   0.685   0.685  0.626  0.625   0.635   63.0   66.0   5.0   0.0
40/40      0.0     -3.0     3.0   0.050   0.708   0.708  0.773  0.772   0.893   63.0   66.0   5.0   0.5
40/40      0.0     -3.0     3.0   0.050   0.747   0.746  0.940  0.939   0.996   63.0   66.0   5.0   0.9

Alpha Comparison for Testing Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|

             H1    Lower   Upper                         Trim.   Trim.   Mann
           Diff   Equiv.  Equiv.  Target  T-Test  Welch T-Test  Welch  Whit'y
N1/N2  (Diff1)    Limit   Limit   Alpha   Alpha   Alpha  Alpha  Alpha   Alpha     M0     M1     S     G
40/40      0.0     -3.0     3.0   0.050   0.048   0.048  0.049  0.049   0.051   63.0   66.0   5.0   0.0
40/40      0.0     -3.0     3.0   0.050   0.043   0.043  0.043  0.042   0.047   63.0   66.0   5.0   0.5
40/40      0.0     -3.0     3.0   0.050   0.055   0.055  0.058  0.057   0.056   63.0   66.0   5.0   0.9

Pool Size: 10000. Simulations: 2000. Run Time: 3.01 minutes. Percent Trimmed: 10.
```



We see that as the degree of skewness is increased, the power of the t-test increases slightly, but the powers of the trimmed and nonparametric tests improve dramatically. The significance levels do not appear to be adversely impacted.

Example 5 – Validation using Machin

Machin *et al.* (1997) page 107 present an example of determining the sample size for a parallel-group design in which the reference mean is 96, the treatment mean is 94, the standard deviation is 8, the limits are plus or minus 5, the power is 80%, and the significance level is 0.05. They calculate the sample size to be 88. It is important to note that Machin *et al.* use an approximation, so their results cannot be expected to exactly match those of ***PASS***.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.80**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Group 1 Distribution | H0	**N(M0 S)**
Group 2 Distribution | H0	**N(91 S)**
Group 1 Distribution | H1	**N(M0 S)**
Group 2 Distribution | H1	**N(94 S)**
M0 (Mean|H0)	**96**
M1 (Mean|H1)	**1**
S	**8**

Output

Click the Run button to perform the calculations and generate the following output.

```
H0 Dist's: Normal(M0 S) & Normal(91 S)
H1 Dist's: Normal(M0 S) & Normal(94 S)
Test Statistic: T-Test

                           Lower    Upper
                    H1    Equiv.   Equiv.   Target   Actual
Power    N1/N2    Diff1    Limit    Limit    Alpha    Alpha     Beta       M0        S
0.807    87/87      2.0     -5.0      5.0    0.050    0.049    0.193     96.0      8.0
(0.017)  [0.790   0.824]                             (0.009)  [0.039    0.058]

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 60.05 seconds.
```

The sample size of 87 per group is reasonably close to the analytic answer of 88.

Chapter 470

Equivalence Tests for Two Means using Ratios

Introduction

This procedure calculates power and sample size of statistical tests for *equivalence* tests from parallel-group design with two groups. This routine deals with the case in which the statistical hypotheses are expressed in terms of mean ratios rather than mean differences.

The details of testing the equivalence of two treatments using a parallel-group design are given in the chapter entitled "Equivalence Tests for Two Means using Differences" and will not be repeated here. If the logarithms of the responses can be assumed to follow a normal distribution, hypotheses about equivalence in terms of the ratio can be transformed into hypotheses about the difference. The details of this analysis are given in Julious (2004).

Equivalence Testing Using Ratios

It will be convenient to adopt the following specialize notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ϕ_L, ϕ_U	RL, RU	*Margin of equivalence.* These limits define an interval of the ratio of the means in which their difference is so small that it may be ignored.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their ratio is needed for power and sample size calculations.

The null hypothesis of non-equivalence is

$$H_0: \phi \le \phi_L \text{ or } \phi \ge \phi_U \textit{ where } \phi_L < 1, \phi_U > 1.$$

and the alternative hypothesis of equivalence is

$$H_1: \phi_L < \phi < \phi_U$$

Log-Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as scale-less percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of ratios.
2. Transform these into hypotheses about differences by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the null hypothesis are as follows.

$$\begin{aligned} &\phi_L \le \phi \le \phi_U \\ &\Rightarrow \phi_L \le \left\{ \frac{\mu_T}{\mu_R} \right\} \le \phi_U \\ &\Rightarrow \ln(\phi_L) \le \{\ln(\mu_T) - \ln(\mu_R)\} \le \ln(\phi_U) \end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

When performing an equivalence test on the difference between means, the usual procedure is to set the equivalence limits symmetrically above and below zero. Thus the equivalence limits will be plus or minus an appropriate amount. The common practice is to do the same when the data are being analyzed on the log scale. However, when symmetric limits are set on the log scale, they do not translate to symmetric limits on the original scale. Instead, they translate to limits that are the inverses of each other.

Perhaps these concepts can best be understood by considering an example. Suppose the researchers have determined that the lower equivalence limit should be 80% on the original scale. Since they are planning to use a log scale for their analysis, they transform this limit to the log scale by taking the logarithm of 0.80. The result is -0.223144. Wanting symmetric limits, they set the upper equivalence limit to 0.223144. Exponentiating this value, they find that exp(0.223144) = 1.25. Note that 1/(0.80) = 1.25. Thus, the limits on the original scale are 80% and 125%, not 80% and 120%.

Using this procedure, appropriate equivalence limits for the ratio of two means can be easily determined. Here are a few sets of equivalence limits.

Specified Percent Change	Lower Limit Original Scale	Upper Limit Original Scale	Lower Limit Log Scale	Upper Limit Log Scale
-25%	75.0%	133.3%	-0.287682	0.287682
+25%	80.0%	125.0%	-0.223144	0.223144
-20%	80.0%	125.0%	-0.223144	0.223144
+20%	83.3%	120.0%	-0.182322	0.182322
-10%	90.0%	111.1%	-0.105361	0.105361
+10%	90.9%	110.0%	-0.095310	0.095310

Note that negative percent-change values specify the lower limit first, while positive percent-change values specify the upper limit first. After the first limit is found, the other limit is calculated as its inverse.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter can be used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable *X* is the logarithm of the original variable *Y*. That is, *X* = ln(*Y*) and *Y* = exp(*X*). Label the mean and variance of *X* as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of *Y* as μ_Y and σ_Y^2, respectively. If *X* is normally distributed, then *Y* is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be expressed as

$$COV_Y = \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y}$$

$$= \sqrt{e^{\sigma_X^2} - 1}$$

Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2 + 1 \right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

One final note: for parallel-group designs, σ_X^2 equals σ_d^2, the average variance used in the t-test of the logged data.

Thus, the hypotheses can be stated in the original (Y) scale and then the power can be analyzed in the transformed (X) scale.

Power Calculation

As is shown above, the hypotheses can be stated in the original (Y) scale using ratios or the logged (X) scale using differences. Either way, the power and sample size calculations are made using the formulas for testing the equivalence of the difference in two means. These formulas are presented another chapter and are not duplicated here.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either Beta for a power analysis or *N1* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of nonequivalent means when in fact the means are equivalent.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of non-equivalent means when in fact the means are nonequivalent.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

You can enter a range of values such as *0.05*, *0.10*, *0.15* or *0.05 to 0.15 by 0.01*.

Sample Size

N1 (Sample Size Reference Group)

Enter a value (or range of values) for the sample size of group 1(the reference group). Note that these values are ignored when you are solving for N1. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Treatment Group)

Enter a value (or range of values) for the sample size of group 2 (the treatment group) or enter *Use R* to base N2 on the value of N1. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and [*Y*] means take the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when N2 is set to *Use R*.

When used, N2 is calculated from N1 using the formula: N2 = [R(N1)] where [Y] is the next integer greater than or equal to Y. Note that setting R = 1.0 forces N2 = N1.

Effect Size – Equivalence Limits

RU (Upper Equivalence Limit)

Enter the upper equivalence limit for the ratio of the two means. When the ratio of the means is between this value and RL, the two means are said to be equivalent. The value must be greater than one. A popular choice is 1.25. Note that this value is not a percentage.

If you enter *1/RL*, then 1/RL will be calculated and used here. This choice is commonly used because RL and 1/RL give limits that are of equal magnitude on the log scale.

RL (Lower Equivalence Limit)

Enter the lower equivalence limit for the ratio of the two means. When the ratio of the means is between this value and RU, the two means are said to be equivalent. The value must be less than one. A popular choice is 0.80. Note that this value is not a percentage.

If you enter *1/RU*, then 1/RU will be calculated and used here. This choice is commonly used because RU and 1/RU give limits that are of equal magnitude on the log scale.

Effect Size – True Ratio

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Usually, the ratio will be assumed to be one. However, some authors recommend calculating the power using a ratio of 1.05 since this will require a larger sample size.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not logged) scale. This value must be determined from past experience or from a pilot study. It is most easily calculated from the within mean-square error of the analysis of variance of the logged data using the relationship

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\,.$$

If prior studies used a t-test to analyze the logged data, you will not have a direct estimate of $\hat{\sigma}_w^2$. However, the two variances, σ_d^2 and σ_w^2, are functionally related. The relationship between these quantities is $\sigma_d^2 = 2\sigma_w^2$.

Example 1 – Finding Power

A company has developed a generic drug for treating rheumatism and wants to show that it is equivalent to the standard drug. A parallel-group design will be used to test the equivalence of the two drugs.

Researchers have decided to set the lower limit of equivalence at 0.80. Past experience leads the researchers to set the COV to 1.50. The significance level is 0.05. The power will be computed assuming that the true ratio is either 1.00 or 1.05. Sample sizes between 50 and 550 will be included in the analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**50 to 550 by 100**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1**
RU (Upper Equivalence Limit)	**1/RL**
RL (Lower Equivalence Limit)	**0.80**
R1 (True Ratio)	**1.0 1.05**
COV (Coefficient of Variation)	**1.50**

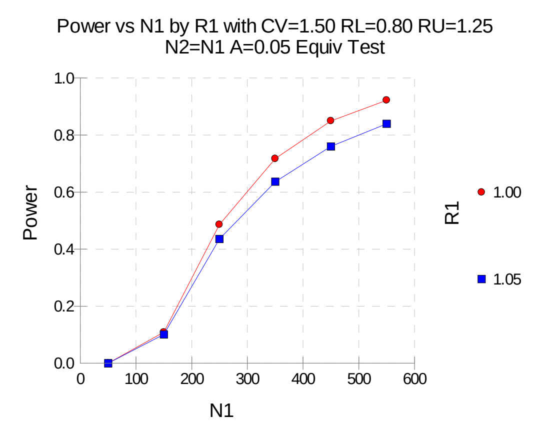
Annotated Output

Click the Run button to perform the calculations and generate the following output.

```
Numeric Results for Testing Equivalence Using a Parallel-Group Design

         Reference   Treatment
             Group       Group       Lower       Upper                Coefficient
            Sample      Sample      Equiv.      Equiv.        True             of
              Size        Size       Limit       Limit       Ratio      Variation
Power         (N1)        (N2)        (RL)        (RU)        (R1)          (COV)      Alpha        Beta
0.0000          50          50        0.80        1.25        1.00           1.50     0.0500      1.0000
0.1049         150         150        0.80        1.25        1.00           1.50     0.0500      0.8951
0.4843         250         250        0.80        1.25        1.00           1.50     0.0500      0.5157
0.7170         350         350        0.80        1.25        1.00           1.50     0.0500      0.2830
0.8494         450         450        0.80        1.25        1.00           1.50     0.0500      0.1506
0.9221         550         550        0.80        1.25        1.00           1.50     0.0500      0.0779
0.0000          50          50        0.80        1.25        1.05           1.50     0.0500      1.0000
0.1010         150         150        0.80        1.25        1.05           1.50     0.0500      0.8990
0.4360         250         250        0.80        1.25        1.05           1.50     0.0500      0.5640
0.6366         350         350        0.80        1.25        1.05           1.50     0.0500      0.3634
0.7602         450         450        0.80        1.25        1.05           1.50     0.0500      0.2398
0.8396         550         550        0.80        1.25        1.05           1.50     0.0500      0.1604

Report Definitions
Power is the probability of rejecting non-equivalence when they are equivalent.
N1 is the number of subjects in the first group.
N2 is the number of subjects in the second group.
RU & RL are the maximum allowable ratios that result in equivalence.
R1 is the ratio of the means at which the power is computed.
COV is the coefficient of variation on the original scale.
Alpha is the probability of rejecting non-equivalence when the means are non-equivalent.
Beta is the probability of accepting non-equivalence when the means are equivalent.

Summary Statements
An equivalence test of means using two one-sided tests on data from a parallel-group design
with sample sizes of 50 in the reference group and 50 in the treatment group achieves 0% power
at a 5% significance level when the true ratio of the means is 1.00, the coefficient of
variation on the original, unlogged scale is 1.50, and the equivalence limits of the mean ratio
are 0.80 and 1.25.
```

This report shows the power for the indicated scenarios.

Plot Section



This plot shows the power versus the sample size.

Example 2 – Validation using Julious

Julious (2004) page 1971 presents an example of determining the sample size for a parallel-group design in which the actual ratio is 1.0, the coefficient of variation is 0.80, the equivalence limits are 0.80 and 1.25, the power is 90%, and the significance level is 0.05. He calculates the per group sample size to be 216.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1**
RU (Upper Equivalence Limit)	**1/RL**
RL (Lower Equivalence Limit)	**0.80**
R1 (True Ratio)	**1.0**
COV (Coefficient of Variation)	**0.80**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Parallel-Group Design

        Reference  Treatment
            Group      Group                                 Coefficient
           Sample     Sample    Lower    Upper                        of
             Size       Size   Equiv.   Equiv.      True       Variation
Power        (N1)       (N2)    Limit    Limit     Ratio           (COV)     Alpha      Beta
0.9004        216        216     (RL)     (RU)      (R1)            0.80    0.0500    0.0996
```

PASS has also calculated the per group sample size to be 216, which matches Julious's result.

Chapter 471

Confidence Intervals for the Difference Between Two Means

Introduction

This procedure calculates the sample size necessary to achieve a specified distance from the difference in sample means to the confidence limit(s) at a stated confidence level for a confidence interval about the difference in means when the underlying data distribution is normal.

Caution: This procedure assumes that the standard deviations of the future samples will be the same as the standard deviations that are specified. If the standard deviation to be used in the procedure is estimated from a previous sample or represents the population standard deviation, the Confidence Intervals for the Difference between Two Means with Tolerance Probability procedure should be considered. That procedure controls the probability that the distance from the difference in means to the confidence limits will be less than or equal to the value specified.

Technical Details

There are two formulas for calculating a confidence interval for the difference between two population means. The different formulas are based on whether the standard deviations are assumed to be equal or unequal.

For each of the cases below, let the means of the two populations be represented by μ_1 and μ_2, and let the standard deviations of the two populations be represented as σ_1 and σ_2.

Case 1 – Standard Deviations Assumed Equal

When $\sigma_1 = \sigma_2 = \sigma$ are unknown, the appropriate two-sided confidence interval for μ_1 - μ_2 is

$$\overline{X}_1 - \overline{X}_2 \pm t_{1-\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Upper and lower one-sided confidence intervals can be obtained by replacing $\alpha/2$ with α.

The required sample size for a given precision, D, can be found by solving the following equation iteratively

$$D = t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

This equation can be used to solve for D or n_1 or n_2 based on the values of the remaining parameters.

Case 2 – Standard Deviations Assumed Unequal

When $\sigma_1 \neq \sigma_2$ are unknown, the appropriate two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\overline{X}_1 - \overline{X}_2 \pm t_{1-\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$$

In this case t is an approximate t and the method is known as the Welch-Satterthwaite method. Upper and lower one-sided confidence intervals can be obtained by replacing $\alpha/2$ with α.

The required sample size for a given precision, D, can be found by solving the following equation iteratively

$$D = t_{1-\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This equation can be used to solve for D or n_1 or n_2 based on the values of the remaining parameters.

Confidence Level

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of n_1 and n_2 items are drawn from populations using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is $1 - \alpha$.

Notice that is a long term statement about many, many samples.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters.

Confidence

Confidence Level (1 – Alpha)

The confidence level, 1 – α, has the following interpretation. If thousands of samples of n1 and n2 items are drawn from populations using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is 1 – α.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90, 0.95* or *0.90 to 0.99 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2,* select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2* = [*R*(*N1*)] where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the interval to be used will be a one-sided or a two-sided confidence interval.

Precision

Distance from Mean Difference to Limit(s)

This is the distance from the confidence limit(s) to the difference in means. For two-sided intervals, it is also known as the precision, half-width, or margin of error.

You can enter a single value or a list of values. The value(s) must be greater than zero.

Standard Deviations

S1 and S2 (Standard Deviations)

Enter an estimate of the standard deviation of group 1 or 2. The standard deviation must be a positive number.

Caution: The sample size estimates for this procedure assume that the standard deviation that is achieved when the confidence interval is produced is the same as the standard deviation entered here.

Press the 'Standard Deviation Estimator' button to obtain help on estimating the standard deviation.

You can enter a range of values such as *1, 2, 3* or *1 to 10 by 1*.

Standard Deviation Equality Assumption

Specify whether the standard deviations are assumed to be the same or different. The choice will determine which of the two common confidence interval formulas for estimating the difference in population means will be used.

- **Assume S1 and S2 are Unequal**

  When the standard deviations are assumed to be unequal, the variances are not pooled and an approximate method is used for the confidence interval formula. This approximate method is sometimes called the Welch-Satterthwaite method.

- **Assume S1 and S2 are Equal**

  When the standard deviations are assumed to be equal, the pooled variance formula is used in the calculation of the confidence interval. The degrees of freedom are N1 + N2 – 2.

  Recommendation: Because the standard deviations of two populations are rarely equal, it is recommended that the standard deviations are assumed to be unequal. The Welch-Satterthwaite confidence interval calculation is generally accepted and commonly used.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Sample Size

Suppose a study is planned in which the researcher wishes to construct a two-sided 95% confidence interval for the difference between two population means such that the width of the interval is no wider than 20 units. The confidence level is set at 0.95, but 0.99 is included for comparative purposes. The standard deviation estimates, based on the range of data values, are 32 for Population 1 and 38 for Population 2. Instead of examining only the interval half-width of 10, a series of half-widths from 5 to 15 will also be considered.

The goal is to determine the necessary sample size for each group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Difference Between Two Means** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Confidence Level	**0.95 0.99**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Interval Type	**Two-Sided**
Distance from Mean Diff to Limit(s)	**5 to 15 by 1**
S1 (Standard Deviation Group 1)	**32**
S2 (Standard Deviation Group 2)	**38**
SD Equality Assumption	**Assume S1 and S2 are Unequal**

Annotated Output

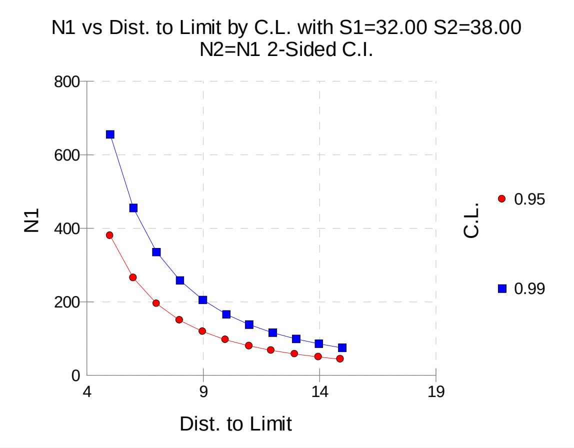
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals for the Difference in Means
The standard deviations are assumed to be Unknown and Unequal.

                                                  Target       Actual
                                                Dist from    Dist from
Confidence                          Allocation  Mean Diff    Mean Diff
Level            N1         N2           Ratio  to Limits    to Limits        S1         S2
0.95            380        380           1.000      5.000        4.995     32.00      38.00
0.95            265        265           1.000      6.000        5.995     32.00      38.00
0.95            195        195           1.000      7.000        6.995     32.00      38.00
0.95            150        150           1.000      8.000        7.984     32.00      38.00
0.95            119        119           1.000      9.000        8.973     32.00      38.00
0.95             97         97           1.000     10.000        9.951     32.00      38.00
0.95             80         80           1.000     11.000       10.973     32.00      38.00
0.95             68         68           1.000     12.000       11.918     32.00      38.00
0.95             58         58           1.000     13.000       12.926     32.00      38.00
0.95             50         50           1.000     14.000       13.947     32.00      38.00
0.95             44         44           1.000     15.000       14.895     32.00      38.00
0.99            655        655           1.000      5.000        5.000     32.00      38.00
0.99            455        455           1.000      6.000        5.999     32.00      38.00
0.99            335        335           1.000      7.000        6.991     32.00      38.00
0.99            258        258           1.000      8.000        7.997     32.00      38.00
0.99            205        205           1.000      9.000        8.981     32.00      38.00
0.99            166        166           1.000     10.000        9.991     32.00      38.00
0.99            138        138           1.000     11.000       10.972     32.00      38.00
0.99            116        116           1.000     12.000       11.983     32.00      38.00
0.99             99         99           1.000     13.000       12.991     32.00      38.00
0.99             86         86           1.000     14.000       13.960     32.00      38.00
0.99             75         75           1.000     15.000       14.975     32.00      38.00

References
Ostle, B. and Malone, L.C. 1988. Statistics in Research. Iowa State University Press. Ames, Iowa.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions
Confidence level is the proportion of confidence intervals (constructed with this same confidence level,
   sample size, etc.) that would contain the true difference in population means.
N1 and N2 are the sample sizes drawn from the two populations.
Allocation Ratio is the ratio of the sample sizes, N2/N1.
Dist from Mean Diff to Limit is the distance from the confidence limit(s) to the difference in sample means.
   For two-sided intervals, it is also know as the precision, half-width, or margin of error.
Target Dist from Mean Diff to Limit is the value of the distance that is entered into the procedure.
Actual Dist from Mean Diff to Limit is the value of the distance that is obtained from the procedure.
S1 and S2 are the standard deviations upon which the distance from mean difference to limit calculations are
   based.

Summary Statements
Group sample sizes of 380 and 380 produce a two-sided 95% confidence interval with a distance
from the difference in means to the limits that is equal to 4.995 when the estimated standard
deviations are 32.00 and 38.00.
```

This report shows the calculated sample size for each of the scenarios.

Plots Section



This plot shows the sample size of each group versus the precision for the two confidence levels.

Example 2 – Validation using Ostle and Malone

Ostle and Malone (1988) page 150 give an example of a precision calculation for a confidence interval for the difference between two means when the confidence level is 95%, the two standard deviations are 6.2185 and 16.06767, and the sample sizes are 7 and 6. The precision is 13.433 (when df = 6.257, not 6).

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Difference Between Two Means** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Confidence Level	**0.90**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**6**
R (Sample Allocation Ratio)	***Ignored***
Interval Type	**Two-Sided**
Distance from Mean Diff to Limit(s)	**13.433**
S1 (Standard Deviation Group 1)	**6.2185**
S2 (Standard Deviation Group 2)	**16.06767**
SD Equality Assumption	**Assume S1 and S2 are Unequal**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals for the Difference in Means
The standard deviations are assumed to be Unknown and Unequal.

                                                     Target      Actual
                                                  Dist from   Dist from
Confidence                           Allocation   Mean Diff   Mean Diff
Level          N1         N2              Ratio   to Limits   to Limits        S1        S2
0.90            7          6              0.857      13.433      13.433      6.22     16.07
```

PASS also calculated the sample size in Group 1 to be 7.

Example 3 – Validation using Zar

Zar (1984) page 132 gives an example of a precision calculation for a confidence interval for the difference between two means when the confidence level is 95%, the pooled standard deviation estimate is 0.7206, and the sample sizes are 6 and 7. The precision is 0.88.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Difference Between Two Means** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Distance from Mean Difference to Limit**
Confidence Level	**0.95**
N1 (Sample Size Group 1)	**6**
N2 (Sample Size Group 2)	**7**
R (Sample Allocation Ratio)	***Ignored***
Interval Type	**Two-Sided**
Distance from Mean Diff to Limit(s)	*Ignored since this is the Find setting*
S1 (Standard Deviation Group 1)	**0.7206**
S2 (Standard Deviation Group 2)	**S1**
SD Equality Assumption	**Assume S1 and S2 are Equal**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals for the Difference in Means
The standard deviations are assumed to be Unknown and Equal.

                                                      Target        Actual
                                                   Dist from     Dist from
Confidence                           Allocation    Mean Diff     Mean Diff
Level          N1          N2             Ratio    to Limits     to Limits       S1        S2
0.95            6           7             1.167                      0.882     0.72      0.72
```

PASS also calculated the precision to be 0.88.

Chapter 472

Confidence Intervals for the Difference Between Two Means with Tolerance Probability

Introduction

This procedure calculates the sample size necessary to achieve a specified distance from the difference in sample means to the confidence limit(s) with a given tolerance probability at a stated confidence level for a confidence interval about the difference in means when the underlying data distribution is normal.

Sample sizes are calculated only for the case where the standard deviations are assumed to be equal, wherein the pooled standard deviation formula is used.

Technical Details

Let the means of the two populations be represented by μ_1 and μ_2, and let the standard deviations of the two populations be represented as σ_1 and σ_2.

When $\sigma_1 = \sigma_2 = \sigma$ are unknown, the appropriate two-sided confidence interval for μ_1 - μ_2 is

$$\overline{X}_1 - \overline{X}_2 \pm t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Upper and lower one-sided confidence intervals can be obtained by replacing $\alpha / 2$ with α.

The required sample size for a given precision, D, can be found by solving the following equation iteratively

$$D = t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

This equation can be used to solve for D or n_1 or n_2 based on the values of the remaining parameters.

There is an additional subtlety that arises when the standard deviation is to be chosen for estimating sample size. The sample sizes determined from the formula above produce confidence intervals with the specified widths only when the future samples have a pooled standard deviation that is no greater than the value specified.

As an example, suppose that 15 individuals are sampled from each population in a pilot study, and a pooled standard deviation estimate of 5.4 is obtained from the sample. The purpose of a later study is to estimate the difference in means within 10 units. Suppose further that the sample size needed is calculated to be 62 per group using the formula above with 5.4 as the estimate for the pooled standard deviation. The samples of size 62 are then obtained from each population, but the pooled standard deviation turns out to be 6.3 rather than 5.4. The confidence interval is computed and the distance from the difference in means to the confidence limits is greater than 10 units.

This example illustrates the need for an adjustment to adjust the sample size such that the distance from the difference in means to the confidence limits will be below the specified value with known probability.

Such an adjustment for situations where a previous sample is used to estimate the standard deviation is derived by Harris, Horvitz, and Mood (1948) and discussed in Zar (1984). The adjustment is

$$D = t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{F_{1-\gamma; n_1+n_2-2, m_1+m_2-2}}$$

where $1 - \gamma$ is the probability that the distance from the difference in means to the confidence limit(s) will be below the specified value, and m_1 and m_2 are the sample sizes in the previous samples that were used to estimate the pooled standard deviation.

The corresponding adjustment when no previous sample is available is discussed in Kupper and Hafner (1989). The adjustment in this case is

$$D = t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{\chi^2_{1-\gamma, n_1+n_2-2}}{n_1 + n_2 - 2}}$$

where, again, $1 - \gamma$ is the probability that the distance from the difference in means to the confidence limit(s) will be below the specified value.

Each of these adjustments accounts for the variability in a future estimate of the pooled standard deviation. In the first adjustment formula (Harris, Horvitz, and Mood, 1948), the distribution of the pooled standard deviation is based on the estimate from previous samples. In the second adjustment formula, the distribution of the pooled standard deviation is based on a specified value that is assumed to be the population pooled standard deviation.

Confidence Level

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of n_1 and n_2 items are drawn from populations using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is $1 - \alpha$.

Notice that is a long term statement about many, many samples.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters.

Confidence and Tolerance

Confidence Level (1 – Alpha)

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of n_1 and n_2 items are drawn from populations using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is $1 - \alpha$.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90 ,0.95* or *0.90 to 0.99 by 0.01*.

Tolerance Probability

This is the probability that a future interval with sample sizes N1 and N2 and the specified confidence level will have a distance from the difference in means to the limit(s) that is less than or equal to the distance specified.

If a tolerance probability is not used, as in the 'Confidence Intervals for the Difference between Two Means' procedure, the sample size is calculated for the expected distance from the difference in means to the limit(s), which assumes that the future standard deviation will also be the one specified.

Using a tolerance probability implies that the standard deviation of the future sample will not be known in advance, and therefore, an adjustment is made to the sample size formula to account for the variability in the standard deviation. Use of a tolerance probability is similar to using an upper bound for the standard deviation in the 'Confidence Intervals for the Difference between Two Means' procedure.

The range of values that can be entered here is values between 0 and 1.

You can enter a range of values such as *.70 .80 .90* or *.70 to .95 by .05*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: $N2 = [R(N1)]$ where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the interval to be used will be a one-sided or a two-sided confidence interval.

Precision

Distance from Mean Difference to Limit(s)

This is the distance from the confidence limit(s) to the difference in means. For two-sided intervals, it is also known as the precision, half-width, or margin of error.

You can enter a single value or a list of values. The value(s) must be greater than zero.

Pooled Standard Deviation

Standard Deviation Source

This procedure permits two sources for estimates of the pooled standard deviation:

- **S is a Population Standard Deviation**

  This option should be selected if there are no previous samples that can be used to obtain an estimate of the pooled standard deviation. In this case, the algorithm assumes that the future sample obtained will be from a population with standard deviation S.

- **S from a Previous Sample**

  This option should be selected if the estimate of the pooled standard deviation is obtained from previous random samples from the same distributions as those to be sampled. The total sample size of the previous samples must also be entered under 'Total Sample Size of Previous Sample'.

Pooled Standard Deviation – S is a Population Standard Deviation

S (Standard Deviation)

Enter an estimate of the pooled standard deviation (must be positive). In this case, the algorithm assumes that future samples obtained will be from a population with pooled standard deviation S.

One common method for estimating the standard deviation is the range divided by 4, 5, or 6.

You can enter a range of values such as *1 2 3* or *1 to 10 by 1*.

Press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Pooled Standard Deviation – S from a Previous Sample

S (Standard Deviation)

Enter an estimate of the pooled standard deviation from a previous (or pilot) study. This value must be positive.

A range of values may be entered.

Press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Total Sample Size of Previous Sample

Enter the total sample size that was used to estimate the pooled standard deviation entered in S (SD Estimated from a Previous Sample). The total sample size should be the total of the two sample sizes (m_1 + m_2) that were used to estimate the pooled standard deviation.

If the previous sample used for the estimate of the pooled standard deviation is a single sample rather than two samples, enter the sample size of the previous sample plus one.

This value is entered only when 'Standard Deviation Source:' is set to 'S from a Previous Sample'.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Sample Size

Suppose a study is planned in which the researcher wishes to construct a two-sided 95% confidence interval for the difference between two population means. It is very important that the mean weight is estimated within 10 units. The pooled standard deviation estimate, based on the range of data values, is 25.6. Instead of examining only the interval half-width of 10, a series of half-widths from 5 to 15 will also be considered.

The goal is to determine the sample size necessary to obtain a two-sided confidence interval such that the difference in means is estimated within 10 units. Tolerance probabilities of 0.70 to 0.95 will be examined.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Difference between Two Means with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Two Means with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Confidence Level	**0.95**
Tolerance Probability	**0.70 to 0.95 by 0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Interval Type	**Two-Sided**
Distance from Mean Diff to Limit(s)	**10**
Standard Deviation Source	**S is a Population Standard Deviation**
S	**25.6**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Two-Sided Confidence Intervals for the Difference in Means

Confidence			Allocation	Target Dist from Mean Diff	Actual Dist from Mean Diff	Pooled Standard	Tolerance
Level	N1	N2	Ratio	to Limits	to Limits	Deviation	Probability
0.95	55	55	1.000	10.000	9.994	25.60	0.70
0.95	56	56	1.000	10.000	9.998	25.60	0.75
0.95	58	58	1.000	10.000	9.919	25.60	0.80
0.95	59	59	1.000	10.000	9.951	25.60	0.85
0.95	61	61	1.000	10.000	9.921	25.60	0.90
0.95	63	63	1.000	10.000	9.962	25.60	0.95

References

Kupper, L. L. and Hafner, K. B. 1989. 'How Appropriate are Popular Sample Size Formulas?', The American Statistician, Volume 43, No. 2, pp. 101-105.

Report Definitions

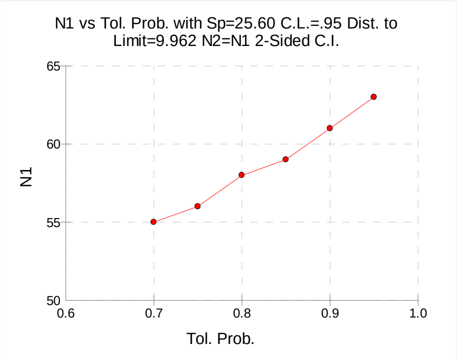
Confidence level is the proportion of confidence intervals (constructed with this same confidence level, sample size, etc.) that would contain the true difference in population means.
N1 and N2 are the sample sizes drawn from the two populations.
Allocation Ratio is the ratio of the sample sizes, N2/N1.
Dist from Mean Diff to Limit(s) is the distance from the confidence limit(s) to the difference in sample means. For two-sided intervals, it is also know as the precision, half-width, or margin of error.
Target Dist from Mean Diff to Limit(s) is the value of the distance that is entered into the procedure.
Actual Dist from Mean Diff to Limit(s) is the value of the distance that is obtained from the procedure.
Pooled Standard Deviation is the standard deviation upon which the distance from mean difference to limit calculations are based.
Tolerance Probability is the probability that a future interval with sample size N and corresponding confidence level will have a distance from the mean to the limit(s) that is less than or equal to the specified distance.

Summary Statements

The probability is 0.70 that group sample sizes of 55 and 55 will produce a two-sided 95% confidence interval with a distance from the difference in means to the limits that is less than or equal to 9.994 if the pooled standard deviation is 25.60.

This report shows the calculated sample size for each of the scenarios.

Plots Section


N1 vs Tol. Prob. with Sp=25.60 C.L.=.95 Dist. to Limit=9.962 N2=N1 2-Sided C.I.

This plot shows the sample size of each group versus the precision for the two confidence levels.

Example 2 – Validation using Zar

Zar (1984) pages 133-134 gives an example of a precision calculation for a confidence interval for the difference between two means when the confidence level is 95%, the pooled standard deviation is 0.720625 from a total sample size of 13, the precision is 0.5, and the tolerance probability is 0.90. The sample size for each group is determined to be 34.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for the Difference between Two Means with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Two Means with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Confidence Level	**0.95**
Tolerance Probability	**0.90**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Interval Type	**Two-Sided**
Distance from Mean Diff to Limit(s)	**0.5**
Standard Deviation Source	**S from a Previous Sample**
S	**0.720625**
Total Sample Size of Previous Sample	**13**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals for the Difference in Means

                                                 Target      Actual
                                              Dist from   Dist from      Pooled
Confidence                       Allocation   Mean Diff   Mean Diff    Standard   Tolerance
Level         N1         N2           Ratio   to Limits   to Limits   Deviation Probability
0.95          34         34           1.000       0.500       0.496        0.72        0.90

Total sample size for estimate of pooled standard deviation from previous samples = 13.
```

PASS also calculated the sample size in each group to be 34.

Chapter 475

Group-Sequential Tests for Two Means

Introduction

Clinical trials are longitudinal. They accumulate data sequentially through time. The participants cannot be enrolled and randomized on the same day. Instead, they are enrolled as they enter the study. It may take several years to enroll enough patients to meet sample size requirements. Because clinical trials are long term studies, it is in the interest of both the participants and the researchers to monitor the accumulating information for early convincing evidence of either harm or benefit. This permits early termination of the trial.

Group sequential methods allow statistical tests to be performed on accumulating data while a phase III clinical trial is ongoing. Statistical theory and practical experience with these designs have shown that making four or five *interim analyses* is almost as effective in detecting large differences between treatment groups as performing a new analysis after each new data value. Besides saving time and resources, such a strategy can reduce the experimental subject's exposure to an inferior treatment and make superior treatments available sooner.

When repeated significance testing occurs on the same data, adjustments have to be made to the hypothesis testing procedure to maintain overall significance and power levels. The landmark paper of Lan & DeMets (1983) provided the theory behind the *alpha spending function* approach to group sequential testing. This paper built upon the earlier work of Armitage, McPherson, & Rowe (1969), Pocock (1977), and O'Brien & Fleming (1979). ***PASS*** implements the methods given in Reboussin, DeMets, Kim, & Lan (1992) to calculate the power and sample sizes of various group sequential designs.

This module calculates sample size and power for group sequential designs used to compare two treatment means. Other modules perform similar analyses for the comparison of proportions and survival functions. The program allows you to vary the number and times of interim tests, the type of alpha spending function, and the test boundaries. It also gives you complete flexibility in solving for power, significance level, sample size, or effect size. The results are displayed in both numeric reports and informative graphics.

Technical Details

Suppose the means of two samples of *N1* and *N2* individuals will be compared at various stages of a trial using the z_k statistic:

$$z_k = \frac{\overline{X}_{1k} - \overline{X}_{2k}}{\sqrt{\frac{s_{1k}^2}{N_{1k}} + \frac{s_{2k}^2}{N_{2k}}}}$$

The subscript *k* indicates that the computations use all data that are available at the time of the k^{th} interim analysis or k^{th} *look* (*k* goes from 1 to *K*). This formula computes the standard *z* test that is appropriate when the variances of the two groups are different. The statistic, z_k, is assumed to be normally distributed.

Spending Functions

Lan and DeMets (1983) introduced alpha spending functions, $\alpha(\tau)$, that determine a set of boundaries $b_1, b_2, \cdots, b_K$ for the sequence of test statistics $z_1, z_2, \cdots, z_K$. These boundaries are the critical values of the sequential hypothesis tests. That is, after each interim test, the trial is continued as long as $|z_k| < b_k$. When $|z_k| \geq b_k$, the hypothesis of equal means is rejected and the trial is stopped early.

The time argument τ either represents the proportion of elapsed time to the maximum duration of the trial or the proportion of the sample that has been collected. When elapsed time is being used it is referred to as *calendar time*. When time is measured in terms of the sample, it is referred to as *information time*. Since it is a proportion, τ can only vary between zero and one.

Alpha spending functions have the characteristics:

$$\alpha(0) = 0$$
$$\alpha(1) = \alpha$$

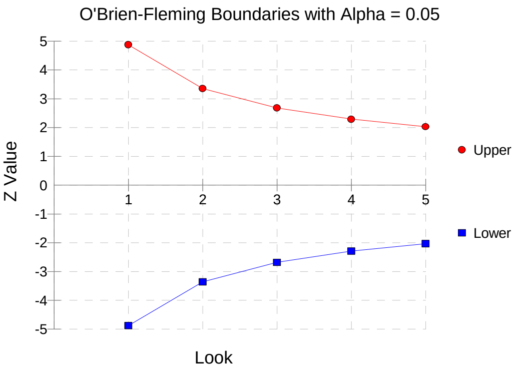
The last characteristic guarantees a fixed α level when the trial is complete. That is,

$$\Pr\left(|z_1| \geq b_1 \ or \ |z_2| \geq b_2 \ or \cdots or \ |z_k| \geq b_k\right) = \alpha(\tau)$$

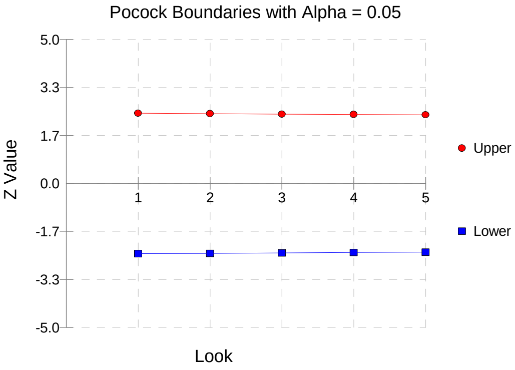
This methodology is very flexible since neither the times nor the number of analyses must be specified in advance. Only the functional form of $\alpha(\tau)$ must be specified.

PASS provides five popular spending functions plus the ability to enter and analyze your own boundaries. These are calculated as follows:

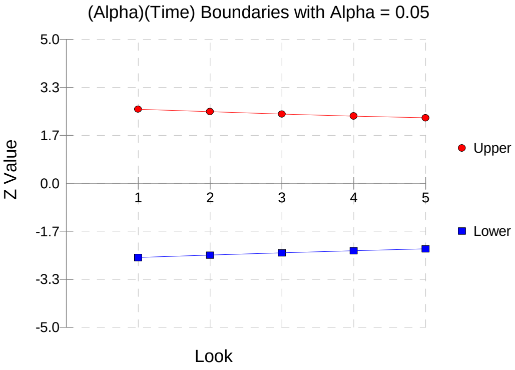
1. O'Brien-Fleming $2-2\Phi\left(\frac{Z_{\alpha/2}}{\sqrt{\tau}}\right)$



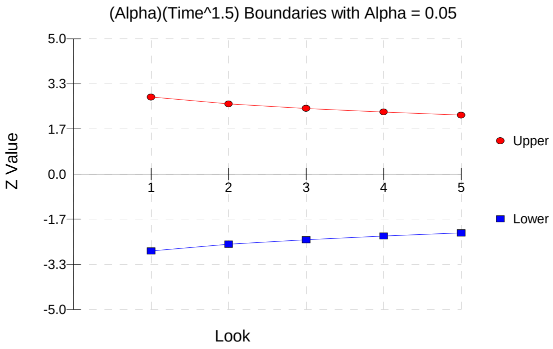
2. Pocock $\alpha\ln\left(1+\left(e-1\right)\tau\right)$



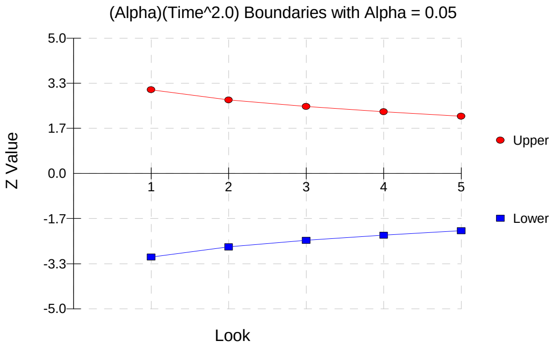
3. Alpha * time $\alpha\tau$



4. Alpha * time^1.5 $\alpha\tau^{3/2}$



5. Alpha * time^2 $\alpha\tau^2$



6. User Supplied

A custom set of boundaries may be entered.

The O'Brien-Fleming boundaries are commonly used because they do not significantly increase the overall sample size and because they are conservative early in the trial. Conservative in the sense that the means must be extremely different before statistical significance is indicated. The Pocock boundaries are nearly equal for all times. The Alpha*t boundaries use equal amounts of alpha when the looks are equally spaced. You can enter your own set of boundaries using the User Supplied option.

Theory

A detailed account of the methodology is contained in Lan & DeMets (1983), DeMets & Lan (1984), Lan & Zucker (1993), and DeMets & Lan (1994). The theoretical basis of the method will be presented here.

Group sequential procedures for interim analysis are based on their equivalence to discrete boundary crossing of a Brownian motion process with drift parameter θ. The test statistics z_k follow the multivariate normal distribution with means $\theta\sqrt{\tau_k}$ and, for $j \le k$, covariances $\sqrt{\tau_k / \tau_j}$. The drift parameter is related to the parameters of the z-test through the equation

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Hence, the algorithm is as follows:

1. Compute boundary values based on a specified spending function and alpha value.
2. Calculate the drift parameter based on those boundary values and a specified power value.
3. Use the drift parameter and estimates of the other parameters in the above equation to calculate the appropriate sample size.

Procedure Tabs

This section describes the options that are unique to this procedure. These are located on the panels associated with the Data, Reports, Bnd Plot Axes, and Options tabs. To find out more about using the other tabs such as Axes/Legend, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with the *z* test such as the means, variances, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Mean1*, *Mean2*, *Alpha*, *Power and Beta*, *N1* or *N2*. Under most situations, you will select either *Power and Beta* or *N1*.

Select *N1* when you want to calculate the sample size needed to achieve a given power and alpha level.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run since power is equal to one minus beta.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact they are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. For this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact they are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Sample Size Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R\ N1]$$

  where *R* is the Sample Allocation Ratio and [Y] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2=[R N1]* where *[Y]* is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Test

Alternative Hypothesis

Specify whether the test is one-sided or two-sided. When a two-sided hypothesis is selected, the value of alpha is halved. Everything else remains the same.

Note that the accepted procedure is to use Two Sided option unless you can justify using a one-sided test.

Effect Size

Mean1

Enter value(s) for the mean of the first group under both hypotheses and the mean of the second group under the null hypothesis of equal means. Note that only the difference between the two means is used in the calculations. You may enter a range of values such as 10,20,30 or 0 to 100 by 25.

If you want to use a single difference rather than the two means, enter the value of the difference as *Mean2* and zero for *Mean1* (or vice versa).

Mean2

Enter value(s) for the mean of the second group under the alternative hypothesis. Note that only the difference between the two means is used in the calculations. You may enter a range of values such as *10,20,30* or *0 to 100 by 25*.

If you want to use a single difference rather than the two means, enter the value of the difference as *Mean2* and zero for Mean1 (or vice versa).

S1 (SD, Group 1)

Enter an estimate of the standard deviation of group 1. The standard deviation must be a positive number. Refer to the chapter on Estimating the Standard Deviation for more information on estimating the standard deviation. Press the *SD* button to obtain a special window designed to help you obtain a realistic value for the standard deviation.

Above all else, remember that the experience of consulting statisticians is that researchers tend to underestimate the standard deviation!

S2 (SD, Group 2)

Enter an estimate of the standard deviation of group 2. The standard deviation must be a positive number. Refer to the chapter on Estimating the Standard Deviation for more information on estimating the standard deviation. Press the *SD* button to obtain a special window designed to help you obtain a realistic value for the standard deviation.

You can enter *S1* here if you want to assume that the standard deviations are equal and use the value entered for *S1*.

Look Details

This box contains the parameters associated with Group Sequential Design such as the type of spending function, the times, and so on.

Number of Looks

This is the number of interim analyses (including the final analysis). For example, a five here means that four interim analyses will be run in addition to the final analysis.

Boundary Truncation

You can truncate the boundary values at a specified value. For example, you might decide that no boundaries should be larger than 4.0. If you want to implement a boundary limit, enter the value here.

If you do not want a boundary limit, enter *None* here.

Spending Function

Specify which alpha spending function to use. The most popular is the O'Brien-Fleming boundary that makes early tests very conservative. Select *User Specified* if you want to enter your own set of boundaries.

Max Time

This is the total running time of the trial. It is used to convert the values in the Times box to fractions. The units (months or years) do not matter, as long as they are consistent with those entered in the Times box.

For example, suppose Max Time = 3 and Times = 1, 2, 3. Interim analyses would be assumed to have occurred at 0.33, 0.67, and 1.00.

Times

Enter a list of time values here at which the interim analyses will occur. These values are scaled according to the value of the Max Time option.

For example, suppose a 48-month trial calls for interim analyses at 12, 24, 36, and 48 months. You could set Max Time to 48 and enter *12,24,36,48* here or you could set Max Time to *1.0* and enter *0.25,0.50,0.75,1.00* here.

The number of times entered here must match the value of the Number of Looks.

- **Equally Spaced**

  If you are planning to conduct the interim analyses at equally spaced points in time, you can enter *Equally Spaced* and the program will generate the appropriate time values for you.

Informations

You can weight the interim analyses on the amount of information obtained at each time point rather than on actual calendar time. If you would like to do this, enter the information amounts here. Usually, these values are the sample sizes obtained up to the time of the analysis.

For example, you might enter *50, 76, 103, 150* to indicate that 50 individuals where included in the first interim analysis, 76 in the second, and so on.

Upper and Lower Boundaries (Spending = User)

If the Spending Function is set to *User Supplied* you can enter a set of lower test boundaries, one for each interim analysis. The lower boundaries should be negative and the upper boundaries should be positive. Typical entries are *4,3,3,3,2* and *4,3,2,2,2*.

- **Symmetric**

  If you only want to enter the upper boundaries and have them copied with a change in sign to the lower boundaries, enter *Symmetric* for the lower boundaries.

Bnd Plot Axes Tab

The Bnd Axes tab, short for Boundary Axes tab, allows the axes of the spending function plots to be set separately from those of the power plots. The options are identical to those of the Axes tab.

Options Tab

The Options tab controls the convergence of the various iterative algorithms used in the calculations.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations to be run before the search for the criterion of interest (Alpha, Beta, etc.) is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank.

Recommended: 500 (or more).

Maximum Iterations (Lan-Demets algorithm)

This is the maximum number of iterations used in the Lan-DeMets algorithm during its search routine. We recommend a value of at least 200.

Tolerance

Probability Tolerance

During the calculation of the probabilities associated with a set of boundary values, probabilities less than this are assumed to be zero.

We suggest a value of 0.00000000001.

Power Tolerance

This is the convergence level for the search for the spending function values that achieve a certain power. Once the iteration changes are less than this amount, convergence is assumed. We suggest a value of 0.0000001.

If the search is too time consuming, you might try increasing this value.

Alpha Tolerance

This is the convergence level for the search for a given alpha value. Once the changes in the computed alpha value are less than this amount, convergence is assumed and iterations stop. We suggest a value of 0.0001.

This option is only used when you are searching for alpha.

If the search is too time consuming, you can try increasing this value.

Example 1 – Finding the Sample Size

A clinical trial is to be conducted over a two-year period to compare the mean response of a new treatment with the current treatment. The current mean is 127 with a standard deviation of 55.88. The health community will be interested in the new treatment if the mean response rate is increased by 20%. So that the sample size requirements for different effect sizes can be compared, it is also of interest to compute the sample size at 10%, 30%, 40%, 50%, 60%, and 70% increases in the response rates.

Testing will be done at the 0.05 significance level and the power should be set to 0.10. A total of four tests are going to be performed on the data as they are obtained. The O'Brien-Fleming boundaries will be used.

Find the necessary sample sizes and test boundaries assuming equal sample sizes per arm and two-sided hypothesis tests.

We could enter these amounts directly into the Group Sequential Means window. Since the base mean is 127, a 20% increase would translate to a new mean response of 127(120/100) = 152.4. The other mean response rates could be computed similarly. However, to make the results more meaningful, we will scale the input by dividing by the current mean. The scaled standard deviation will be 100(55.88)/127 = 44.00. We set Mean1 to zero since we are only interested in the changes in *Mean2*. The values of *Mean2* will then be 10, 20, 30, 40, 50, 60, and 70.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Group-Sequential Tests for Two Means** procedure window by clicking on **Group-Sequential Tests**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	**Ignored**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Alternative Hypothesis	**Two-Sided**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**10 to 70 by 10**
S1 (Standard Deviation Group 1)	**44**
S2 (Standard Deviation Group 2)	**S1**
Number of Looks	**4**
Spending Function	**O'Brien-Fleming**
Times	**Equally Spaced**
Max Time	**2**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

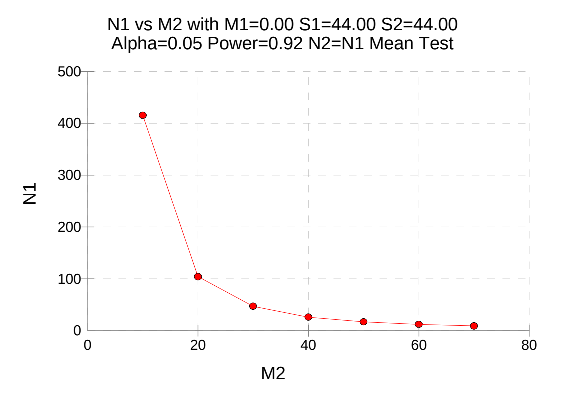
Numeric Results

```
Numeric Results for Two-Sided Hypothesis Test of Means

Power     N1     N2     Alpha     Beta    Mean1   Mean2      S1       S2
0.9005   415    415    0.0500   0.0995     0.00   10.00   44.00    44.00
0.9012   104    104    0.0500   0.0988     0.00   20.00   44.00    44.00
0.9058    47     47    0.0500   0.0942     0.00   30.00   44.00    44.00
0.9012    26     26    0.0500   0.0988     0.00   40.00   44.00    44.00
0.9071    17     17    0.0500   0.0929     0.00   50.00   44.00    44.00
0.9116    12     12    0.0500   0.0884     0.00   60.00   44.00    44.00
0.9170     9      9    0.0500   0.0830     0.00   70.00   44.00    44.00

Report Definitions
Power is the probability of rejecting a false null hypothesis. Power should be close to one.
N1 and N2 are the number of items sampled from groups 1 and 2.
Alpha is the probability of rejecting a true null hypothesis in at least one of the sequential tests.
Beta is the probability of accepting a false null hypothesis at the conclusion of all tests.
Mean1 is the mean of populations 1 and 2 under the null hypothesis of equality.
Mean2 is the mean of population 2 under the alternative hypothesis. The mean of population 1 is unchanged.
S1 and S2 are the population standard deviations of groups 1 and 2.

Summary Statements
Sample sizes of 415 and 415 achieve 90% power to detect a difference of 10.00 between the group
means with standard deviations of 44.00 and 44.00 at a significance level (alpha) of 0.0500
using a two-sided z-test. These results assume that 4 sequential tests are made using the
O'Brien-Fleming spending function to determine the test boundaries.
```

This report shows the values of each of the parameters, one scenario per row. Note that 104 participants in each arm of the study are required to meet the 90% power requirement when the mean increase is 20%.

The values from this table are in the chart below. Note that this plot actually occurs further down in the report.

Plots Section



This plot shows that a large increase in sample size is necessary to test mean differences below 20%.

Details Section

```
Details when Spending = O'Brien-Fleming, N1 = 415, N2 =415, S1 = 44.00, S2 = 44.00, Diff = -10.00
                        Lower       Upper     Nominal        Inc       Total        Inc       Total
Look     Time           Bndry       Bndry       Alpha      Alpha       Alpha      Power       Power
1        0.50        -4.33263     4.33263    0.000015   0.000015    0.000015   0.003512    0.003512
2        1.00        -2.96311     2.96311    0.003045   0.003036    0.003051   0.254998    0.258510
3        1.50        -2.35902     2.35902    0.018323   0.016248    0.019299   0.427601    0.686111
4        2.00        -2.01406     2.01406    0.044003   0.030701    0.050000   0.214371    0.900483
Drift 3.27383
```

This report shows information about the individual interim tests. One report is generated for each scenario.

Look

These are the sequence numbers of the interim tests.

Time

These are the time points at which the interim tests are conducted. Since the Max Time was set to 2 (for two years), these time values are in years. Hence, the first interim test is at half a year, the second at one year, and so on.

We could have set Max Time to 24 so that the time scale was in months.

Lower and Upper Boundary

These are the test boundaries. If the computed value of the test statistic *z* is between these values, the trial should continue. Otherwise, the trial can be stopped.

Nominal Alpha

This is the value of alpha for these boundaries if they were used for a single, standalone, test. Hence, this is the significance level that must be found for this look in a standard statistical package that does not adjust for multiple looks.

Inc Alpha

This is the amount of alpha that is *spent* by this interim test. It is close to, but not equal to, the value of alpha that would be achieved if only a single test was conducted. For example, if we lookup the third value, 2.35902, in normal probability tables, we find that this corresponds to a (two-sided) alpha of 0.0183. However, the entry is 0.0162. The difference is due to the correction that must be made for multiple tests.

Total Alpha

This is the total amount of alpha that is used up to and including the current test.

Inc Power

These are the amounts that are added to the total power at each interim test. They are often called the exit probabilities because they give the probability that significance is found and the trial is stopped, given the alternative hypothesis.

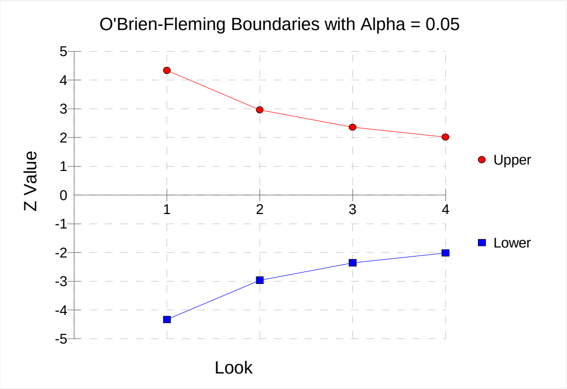
Total Power

These are the cumulative power values. They are also the cumulative exit probabilities. That is, they are the probability that the trial is stopped at or before the corresponding time.

Drift

This is the value of the Brownian motion drift parameter.

Boundary Plots



This plot shows the interim boundaries for each look. This plot shows very dramatically that the results must be extremely significant at early looks, but that they are near the single test boundary (1.96 and -1.96) at the last look.

Example 2 – Finding the Power

A clinical trial is to be conducted over a two-year period to compare the mean response of a new treatment with the current treatment. The current mean is 127 with a standard deviation of 55.88. The health community will be interested in the new treatment if the mean response rate is increased by 20%. The researcher wishes to calculate the power of the design at sample sizes 20, 60, 100, 140, 180, and 220. Testing will be done at the 0.01, 0.05, 0.10 significance levels and the overall power will be set to 0.10. A total of four tests are going to be performed on the data as they are obtained. The O'Brien-Fleming boundaries will be used. Find the power of these sample sizes and test boundaries assuming equal sample sizes per arm and two-sided hypothesis tests.

Proceeding as in Example1, we decide to translate the mean and standard deviation into a percent of mean scale.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Group-Sequential Tests for Two Means** procedure window by clicking on **Group-Sequential Tests**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.01, 0.05, 0.10**
N1 (Sample Size Group 1)	**20 to 220 by 40**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Alternative Hypothesis	**Two-Sided**

Data Tab (continued)

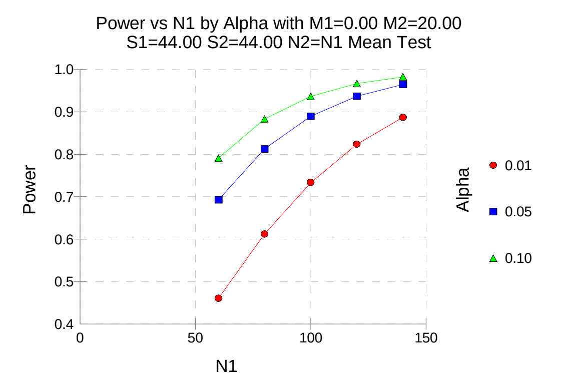
Mean1 (Mean of Group 1)....................... **0**
Mean2 (Mean of Group 2)....................... **20**
S1 (Standard Deviation Group 1)............ **44**
S2 (Standard Deviation Group 2)............ **S1**
Number of Looks.................................... **4**
Spending Function .... **O'Brien-Fleming**
Times.. **Equally Spaced**
Max Time... **2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Hypothesis Test of Means

Power      N1     N2     Alpha     Beta    Mean1    Mean2      S1       S2
0.9005    415    415    0.0500   0.0995    0.00    10.00    44.00    44.00
0.1256     20     20    0.0100   0.8744    0.00    20.00    44.00    44.00
0.4605     60     60    0.0100   0.5395    0.00    20.00    44.00    44.00
0.7335    100    100    0.0100   0.2665    0.00    20.00    44.00    44.00
0.8871    140    140    0.0100   0.1129    0.00    20.00    44.00    44.00
0.9572    180    180    0.0100   0.0428    0.00    20.00    44.00    44.00
0.9851    220    220    0.0100   0.0149    0.00    20.00    44.00    44.00
0.2948     20     20    0.0500   0.7052    0.00    20.00    44.00    44.00
0.6929     60     60    0.0500   0.3071    0.00    20.00    44.00    44.00
0.8897    100    100    0.0500   0.1103    0.00    20.00    44.00    44.00
0.9650    140    140    0.0500   0.0350    0.00    20.00    44.00    44.00
0.9898    180    180    0.0500   0.0102    0.00    20.00    44.00    44.00
0.9972    220    220    0.0500   0.0028    0.00    20.00    44.00    44.00
0.4094     20     20    0.1000   0.5906    0.00    20.00    44.00    44.00
0.7909     60     60    0.1000   0.2091    0.00    20.00    44.00    44.00
0.9368    100    100    0.1000   0.0632    0.00    20.00    44.00    44.00
0.9827    140    140    0.1000   0.0173    0.00    20.00    44.00    44.00
0.9956    180    180    0.1000   0.0044    0.00    20.00    44.00    44.00
0.9989    220    220    0.1000   0.0011    0.00    20.00    44.00    44.00
```



These data show the power for various sample sizes and alphas. It is interesting to note that once the sample size is greater than 150, the value of alpha makes little difference on the value of power.

Example 3 – Effect of Number of Looks

Continuing with examples one and two, it is interesting to determine the impact of the number of looks on power. ***PASS*** allows only one value for the Number of Looks parameter per run, so it will be necessary to run several analyses. To conduct this study, set alpha to 0.05, *N1* to 100, and leave the other parameters as before. Run the analysis with Number of Looks equal to 1, 2, 3, 4, 6, 8, 10, and 20. Record the power for each run.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Group-Sequential Tests for Two Means** procedure window by clicking on **Group-Sequential Tests**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**100**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Alternative Hypothesis	**Two-Sided**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**20**
S1 (Standard Deviation Group 1)	**44**
S2 (Standard Deviation Group 2)	**S1**
Number of Looks	**1 (Also run with 2, 3, 4, 6, 8, 10, and 20)**
Spending Function	**O'Brien-Fleming**
Times	**Equally Spaced**
Max Time	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Hypothesis Test of Means

Power    N1    N2    Alpha    Beta    Mean1   Mean2     S1      S2    Looks
0.8951   100   100   0.0500   0.1049   0.00   20.00   44.00   44.00      1
0.8941   100   100   0.0500   0.1059   0.00   20.00   44.00   44.00      2
0.8916   100   100   0.0500   0.1084   0.00   20.00   44.00   44.00      3
0.8897   100   100   0.0500   0.1103   0.00   20.00   44.00   44.00      4
0.8871   100   100   0.0500   0.1129   0.00   20.00   44.00   44.00      6
0.8856   100   100   0.0500   0.1144   0.00   20.00   44.00   44.00      8
0.8845   100   100   0.0500   0.1155   0.00   20.00   44.00   44.00     10
0.8820   100   100   0.0500   0.1180   0.00   20.00   44.00   44.00     20
```

This analysis shows how little the number of looks impact the power of the design. The power of a study with no interim looks is 0.8951. When twenty interim looks are made, the power falls just 0.0131, to 0.8820—a very small change.

Example 4 – Studying a Boundary Set

Continuing with the previous examples, suppose that you are presented with a set of boundaries and want to find the quality of the design (as measured by alpha and power). This is easy to do with ***PASS***. Suppose that the analysis is to be run with five interim looks at equally spaced time points. The upper boundaries to be studied are 3.5, 3.5, 3.0, 2.5, 2.0. The lower boundaries are symmetric. The analysis would be run as follows.

Setup

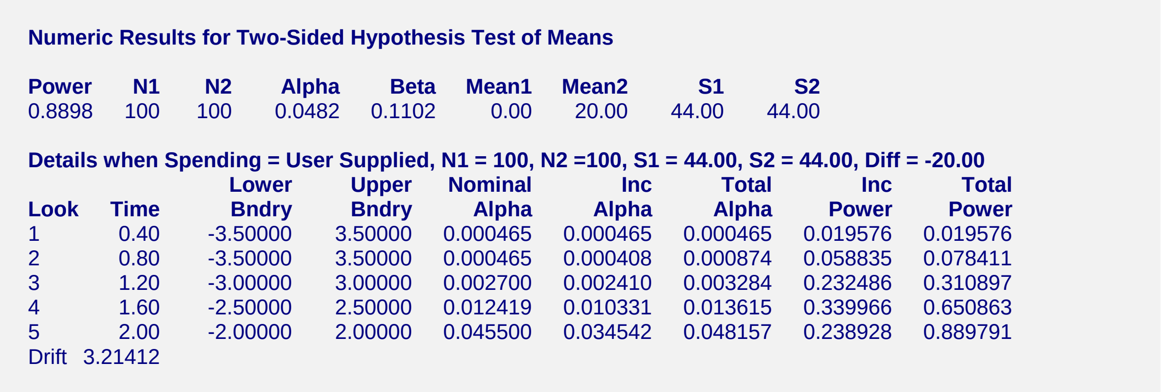
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Group-Sequential Tests for Two Means** procedure window by clicking on **Group-Sequential Tests**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05 (will be calculated from boundaries)**
N1 (Sample Size Group 1)	**100**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Alternative Hypothesis	**Two-Sided**
Mean1 (Mean of Group 1)	**0**
Mean2 (Mean of Group 2)	**20**
S1 (Standard Deviation Group 1)	**44**
S2 (Standard Deviation Group 2)	**S1**
Number of Looks	**5**
Spending Function	**User Supplied**
Max Time	**2**
Times	**Equally Spaced**
Upper Boundaries	**3.5, 3.5, 3.0, 2.5, 2.0**
Lower Boundaries	**Symmetric**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results



```
Numeric Results for Two-Sided Hypothesis Test of Means

Power     N1     N2    Alpha     Beta   Mean1    Mean2      S1       S2
0.8898   100    100   0.0482   0.1102    0.00    20.00   44.00    44.00

Details when Spending = User Supplied, N1 = 100, N2 =100, S1 = 44.00, S2 = 44.00, Diff = -20.00
                   Lower      Upper    Nominal        Inc      Total        Inc      Total
Look    Time       Bndry      Bndry      Alpha      Alpha      Alpha      Power      Power
1       0.40    -3.50000    3.50000   0.000465   0.000465   0.000465   0.019576   0.019576
2       0.80    -3.50000    3.50000   0.000465   0.000408   0.000874   0.058835   0.078411
3       1.20    -3.00000    3.00000   0.002700   0.002410   0.003284   0.232486   0.310897
4       1.60    -2.50000    2.50000   0.012419   0.010331   0.013615   0.339966   0.650863
5       2.00    -2.00000    2.00000   0.045500   0.034542   0.048157   0.238928   0.889791
Drift  3.21412
```

The power for this design is about 0.89. This value depends on both the boundaries and the sample size. The alpha level is 0.048157. This value only depends on the boundaries.

Example 5 – Validation using O’Brien-Fleming Boundaries

Reboussin (1992) presents an example for normally distributed data for a design with two-sided O’Brien-Fleming boundaries, looks = 5, alpha = 0.05, beta = 0.10, *Mean1* = 220, *Mean2* = 200, standard deviation = 30. They compute a drift of 3.28 and a sample size of 48.41 per group. The upper boundaries are: 4.8769, 3.3569, 2.6803, 2.2898, 2.0310.

To test that ***PASS*** provides the same result, enter the following.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Group-Sequential Tests for Two Means** procedure window by clicking on **Group-Sequential Tests**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Alternative Hypothesis	**Two-Sided**
Mean1 (Mean of Group 1)	**220**

Data Tab (continued)

Mean2 (Mean of Group 2).......................	**200**
S1 (Standard Deviation Group 1)............	**30**
S2 (Standard Deviation Group 2)............	**S1**
Number of Looks....................................	**5**
Spending Function	**O'Brien-Fleming**
Max Time...	**1**
Times..	**Equally Spaced**
Upper Boundaries	***Ignored***
Lower Boundaries	***Ignored***

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Hypothesis Test of Means

Power      N1      N2      Alpha       Beta     Mean1     Mean2        S1         S2
0.903623   49      49   0.050000  0.096377    220.00    200.00     30.00      30.00

Details when Spending = O'Brien-Fleming, N1 = 49, N2 =49, S1 = 30.00, S2 = 30.00, Diff = 20.00
                        Lower        Upper      Nominal          Inc         Total          Inc         Total
Look       Time         Bndry        Bndry        Alpha        Alpha         Alpha        Power         Power
1          0.20      -4.87688      4.87688     0.000001     0.000001      0.000001     0.000336      0.000336
2          0.40      -3.35695      3.35695     0.000788     0.000787      0.000788     0.101727      0.102062
3          0.60      -2.68026      2.68026     0.007357     0.006828      0.007616     0.350673      0.452735
4          0.80      -2.28979      2.28979     0.022034     0.016807      0.024424     0.299186      0.751921
5          1.00      -2.03100      2.03100     0.042255     0.025576      0.050000     0.151702      0.903623
Drift   3.29983
```

The slight difference in the power and the drift parameter is attributable to the rounding of the sample size from 48.41 to 49.

Example 6 – Validation with Pocock Boundaries

Reboussin (1992) presents an example for normally distributed data for a design with two-sided Pocock boundaries, looks = 5, alpha = 0.05, beta = 0.10, *Mean1* = 220, *Mean2* = 200, standard deviation = 30. They compute a drift of 3.55 and a sample size of 56.71 per group. The upper boundaries are: 2.4380, 2.4268, 2.4101, 2.3966, and 2.3859.

To test that ***PASS*** provides the same result, enter the following.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Group-Sequential Tests for Two Means** procedure window by clicking on **Group-Sequential Tests**, then **Two Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1**
Power	**0.90**
Alpha	**0.05**
N1 (Sample Size Group 1)	*Ignored since this is the Find setting*
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Alternative Hypothesis	**Two-Sided**
Mean1 (Mean of Group 1)	**220**
Mean2 (Mean of Group 2)	**200**
S1 (Standard Deviation Group 1)	**30**
S2 (Standard Deviation Group 2)	**S1**
Number of Looks	**5**
Spending Function	**Pocock**
Max Time	**1**
Times	**Equally Spaced**
Upper Boundaries	***Ignored***
Lower Boundaries	***Ignored***

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Hypothesis Test of Means

Power        N1    N2     Alpha      Beta  Mean1      Mean2       S1        S2
0.903263     57    57  0.050000  0.096737  220.00     200.00    30.00     30.00

Details when Spending = O'Brien-Fleming, N1 = 49, N2 =49, S1 = 30.00, S2 = 30.00, Diff = 20.00
                          Lower      Upper    Nominal        Inc      Total        Inc      Total
Look       Time           Bndry      Bndry      Alpha      Alpha      Alpha      Power      Power
1          0.20        -2.43798    2.43798   0.014770   0.014770   0.014770   0.198712   0.198712
2          0.40        -2.42677    2.42677   0.015234   0.011387   0.026157   0.260597   0.459308
3          0.60        -2.41014    2.41014   0.015946   0.009269   0.035426   0.214118   0.673426
4          0.80        -2.39658    2.39658   0.016549   0.007816   0.043242   0.143792   0.817218
5          1.00        -2.38591    2.38591   0.017037   0.006758   0.050000   0.086045   0.903263
Drift   3.55903
```

The slight difference in the power and the drift parameter is attributable to the rounding of the sample size from 56.71 to 57.

Chapter 480

Inequality Tests for Two Means in a Cluster-Randomized Design

Introduction

Cluster Randomization refers to the situation in which the means of two groups, made up of *M* clusters of *N* individuals each, are to be tested using a modified *t* test. In this case, the basic experimental unit is a cluster instead of an individual.

Technical Details

Our formulation comes from Donner and Klar (1996). Denote an observation by X_{ijk} where *i* = 1,2 is the group, *j* = 1,2,…,*M* is a cluster in group *i*, and *k* = 1,2,…*N* is an individual in cluster *j* of group *i*. Each cluster mean, $\overline{X}_{ij}$, has a population mean of μ_i and variance

$$Var\left(\overline{X}_i\right) = \left(\frac{\sigma^2}{N}\right)\left[1 + \left(N - 1\right)\rho\right]$$

where σ^2 is the variance of X_{ijk} and ρ is the intracluster correlation coefficient. This correlation made be thought of as the simple correlation between any two observations on the same individual. It may also be thought of as the proportion of total variance in the observations that can be attributed to difference between clusters.

The power for the two-sided, two-sample *t* test using the above formulation is calculated by

$$Power = 1 - P(t \le t_{\alpha/2}, df, \lambda) + P(t \le -t_{\alpha/2}, df, \lambda)$$

where

$df = 2(M-1)$

$$\lambda = \frac{d}{\left[2\left(1+(N-1)\rho\right)/(MN)\right]^{1/2}}$$

$$d = \frac{\left|\mu_1 - \mu_2\right|}{\sigma}$$

Procedure Options

This section describes the options that are unique to this procedure. These are located on the panels associated with the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *D*, *S*, *M*, *N*, *Alpha*, and *Power and Beta*.

Under most situations, you will select either *Power and Beta* to calculate power or *N* to calculate sample size.

Note that the value selected here always appears as the vertical axis on the charts.

The program is set up to evaluate power directly. For the other parameters, a search is made using an iterative procedure until an appropriate value is found.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

If your only interest is in determining the appropriate sample size for a confidence interval, set beta to 0.5.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

M (Number of Clusters)

Enter a value (or range of values) for the number of clusters, *M*, per group.

You may enter a range of values such as *2,4,6* or *2 to 12 by 2*.

N (Individuals Per Cluster)

Enter a value (or range of values) for the number of individuals, *N*, per cluster.

You may enter a range of values such as *100,200,300* or *100 to 300 by 50*.

Effect Size – Mean Difference

D (Difference Between Means)

This is the absolute value of the difference between the two group means. This value, divided by the standard deviation, becomes the effect size.

Effect Size – Standard Deviation

S (Standard Deviation)

Enter a value (or range of values) for the standard deviation. This value is only used as the divisor of the effect size. Hence, if you do not know the standard deviation, you can enter a one here and use effect size units for *D*, the difference.

Remember, this is the standard deviation that occurs when the same individual is measured over and over.

Effect Size – Intracluster Correlation

R (Intracluster Correlation)

Enter a value (or range of values) for the intracluster correlation. This correlation made be thought of as the simple correlation between any two observations on the same individual. It may also be thought of as the proportion of total variance in the observations that can be attributed to difference between clusters.

Although the actual range for this value is from zero to one, typical values range from 0.002 to 0.010.

Test

Alternative Hypothesis

Specify whether the test is one-sided or two-sided. A two-sided hypothesis states that the values are not equal without specifying which is greater. If you do not have any special reason to do otherwise, you should use the two-sided option.

When a two-sided hypothesis is selected, the value of alpha is split in half. Everything else remains the same.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Power

Suppose that a study is to be conducted in which *D* = 0.2; *S* = 1.0; *R* = 0.01; *M* = 6; Alpha = 0.01, 0.05; and *N* = 50 to 300 by 50 and beta is to be calculated.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Cluster-Randomized Design** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Cluster-Randomized Designs**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find parameter*
Alpha	**0.01, 0.05**
M (Number of Clusters)	**6**
N (Individuals Per Cluster)	**50 to 300 by 50**
D (Difference Between Means)	**0.2**
S (Standard Deviation)	**1.0**
R (Intracluster Correlation)	**0.01**
Alternative Hypothesis	**Two-Sided**

Annotated Output

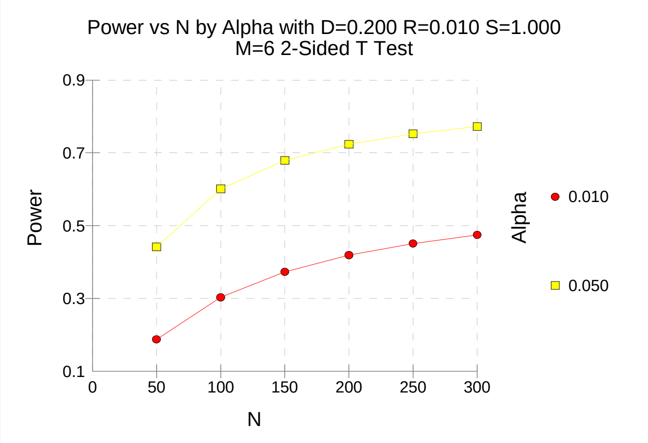
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Test
                     M          N                       R            S
             Number of Individuals            D  Intracluster   Standard
Power         Clusters Per Clusters  Difference  Correlation   Deviation      Alpha       Beta
0.18754              6         50       0.200      0.01000       1.000    0.01000    0.81246
0.44200              6         50       0.200      0.01000       1.000    0.05000    0.55800
0.30320              6        100       0.200      0.01000       1.000    0.01000    0.69680
0.60128              6        100       0.200      0.01000       1.000    0.05000    0.39872
0.37332              6        150       0.200      0.01000       1.000    0.01000    0.62668
0.67912              6        150       0.200      0.01000       1.000    0.05000    0.32088
0.41910              6        200       0.200      0.01000       1.000    0.01000    0.58090
0.72389              6        200       0.200      0.01000       1.000    0.05000    0.27611
0.45101              6        250       0.200      0.01000       1.000    0.01000    0.54899
0.75259              6        250       0.200      0.01000       1.000    0.05000    0.24741
0.47443              6        300       0.200      0.01000       1.000    0.01000    0.52557
0.77242              6        300       0.200      0.01000       1.000    0.05000    0.22758

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
M is the number of clusters per group. There are two groups.
N is the number of individuals per cluster.
D is difference between the group means.
R is intracluster correlation.
S is standard deviation within an individual.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.

Summary Statements
A sample size of 6 clusters per group with 50 individuals per cluster achieves 19% power to
detect a difference of 0.200 between the group means when the standard deviation is 1.000 and
the intracluster correlation is 0.01000 using a two-sided T-test with a significance level of
0.01000.
```

This report shows the power for each of the scenarios.

Plots Section



This plot shows the power versus the cluster size for the two alpha values.

Example 2 – Validation using Donner and Klar

Donner and Klar (1996) page 436 provide a table in which several power values are calculated. When alpha is 0.05, *D* is 0.2, *R* is 0.001, *S* is 1.0, and *M* is 3, they calculate a power of 0.43 for an *N* of 100, 0.79 for an *N* of 300, and 0.91 for an *N* of 500.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a Cluster-Randomized Design** procedure window by clicking on **Means**, then **Two Means**, then **Independent**, then **Inequality Tests**, then **Cluster-Randomized Designs**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find parameter*
Alpha	**0.05**
M (Number of Clusters)	**3**
N (Individuals Per Cluster)	**100 300 500**
D (Difference Between Means)	**0.2**
S (Standard Deviation)	**1.0**
R (Intracluster Correlation)	**0.001**
Alternative Hypothesis	**Two-Sided**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Test
                    M            N                        R           S
            Number of   Individuals         D   Intracluster   Standard
Power        Clusters   Per Clusters  Difference  Correlation  Deviation      Alpha       Beta
0.43008             3           100       0.200      0.00100      1.000    0.05000    0.56992
0.79236             3           300       0.200      0.00100      1.000    0.05000    0.20764
0.90905             3           500       0.200      0.00100      1.000    0.05000    0.09095
```

As you can see, ***PASS*** has calculated the same power values as Donner and Klar (1996).

Chapter 490

Inequality Tests for Paired Means (Simulation)

Introduction

This procedure allows you to study the power and sample size of several statistical tests of the null hypothesis that the difference between two correlated means is equal to a specific value versus the alternative that it is greater than, less than, or not-equal to that value. The paired t-test is commonly used in this situation. Other tests have been developed for the case when the data are not normally distributed. These additional tests include the Wilcoxon signed-ranks test, the sign test, and the computer-intensive bootstrap test.

Paired data may occur because two measurements are made on the same subject or because measurements are made on two subjects that have been matched according to other, often demographic, variables. Hypothesis tests on paired data can be analyzed by considering the differences between the paired items. The distribution of differences is usually symmetric. In fact, the distribution must be symmetric if the individual distributions of the two items are identical. Hence, the paired t-test and the Wilcoxon signed-rank test are appropriate for paired data even when the distributions of the individual items are not normal.

The details of the power analysis of the paired t-test using analytic techniques are presented in another ***PASS*** chapter and they won't be duplicated here. This chapter will only consider power analysis using computer simulation.

Technical Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify the test procedure and the test statistic. This includes the significance level, sample size, and underlying data distributions.

2. Generate a random sample $X_1, X_2, \ldots, X_n$ from the distribution specified by the alternative hypothesis. Calculate the test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. These samples are used to calculate the power of the test. In the case of paired data, the individual values are simulated as the difference between two other random variables. These samples are constructed so that they exhibit a certain amount of correlation.
3. Generate a random sample $Y_1, Y_2, \ldots, Y_n$ from the distribution specified by the null hypothesis. Calculate the test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. These samples are used to calculate the significance-level of the test. In the case of paired data, the individual values are simulated as the difference between two other random variables. These samples are constructed so that they exhibit a certain amount of correlation.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Simulating Paired Distributions

Paired data occur when two observations are correlated. Examples of paired designs are pre – post designs, cross-over designs, and matched pair designs.

In order to simulate paired data, the simulation should mimic the actual data generation process as closely as possible. Since paired data are analyzed by creating the individual difference between each pair, the simulation should also create data as the difference between two variates. Paired data exhibit a correlation between the two variates. As this correlation between the variates increases, the variance of the difference decreases. Thus it is important not only to specify the distributions of the two variates that will be differenced, but to also specify their correlation.

Obtaining paired samples from arbitrary distributions with a set correlation is difficult because the joint, bivariate distribution must be specified and simulated. Rather than specify the bivariate distribution, ***PASS*** requires the specification of the two marginal distributions and the correlation between them.

Monte Carlo samples with given marginal distributions and correlation are generated using the method suggested by Gentle (1998). The method begins by generating a large population of random numbers from the two distributions. Each of these populations is evaluated to determine if their means are within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean.

The next step is to obtain the target correlation. This is accomplished by permuting one of the populations until they have the desired correlation.

The above steps provide a large pool of random numbers that exhibit the desired characteristics. This pool is then sampled at random using the uniform distribution to obtain the random numbers used in the simulation.

This algorithm may be stated as follows.

1. Draw individual samples of size M from the two distributions where M is a large number, usually over 10,000. Adjust these samples so that they have the specified mean and standard deviation. Label these samples A and B. Create an index of the values of A and B according to the order in which they are generated. Thus, the first value of A and the first value of B are indexed as one, the second values of A and B are indexed as two, and so on up to the final set which is indexed as M.
2. Compute the correlation between the two generated variates.
3. If the computed correlation is within a small tolerance (usually less than 0.001) of the specified correlation, go to step 7.
4. Select two indices (I and J) at random using uniform random numbers.
5. Determine what will happen to the correlation if B_I is swapped with B_J. If the swap will result in a correlation that is closer to the target value, swap the indices and proceed to step 6. Otherwise, go to step 4.
6. If the computed correlation is within the desired tolerance of the target correlation, go to step 7. Otherwise, go to step 4.
7. End with a population with the required marginal distributions and correlation.

Now, to complete the simulation, random samples of the designated size are drawn from this population.

Test Statistics

This section describes the test statistics that are available in this procedure. Note that these test statistics are computed on the differences. Thus, when the equation refers to an X value, this X value is assumed to be a difference between two individual variates.

One-Sample t-Test

The one-sample t-test assumes that the data are a simple random sample from a population of normally-distributed values that all have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the t-test proceeds as follow

$$t_{n-1} = \frac{\overline{X} - M0}{s_{\overline{X}}}$$

where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

$$s_{\overline{X}} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}},$$

and *M*0 is the value of the difference hypothesized by the null hypothesis.

The significance of the test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the hypothesis is rejected. Otherwise, no conclusion can be reached.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a popular, nonparametric substitute for the t-test. It assumes that the data follow a symmetric distribution. The test is computed using the following steps.

1. Subtract the hypothesized mean, *D*0, from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks *Sp* and the sum of the negative ranks *Sn*. The test statistic, *W*, is the minimum of *Sp* and *Sn*.
3. Compute the mean and standard deviation of *W* using the formulas

$$\mu_{W_n} = \frac{n(n+1)}{4} \text{ and } \sigma_{W_n} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}$$

   where t_i represents the number of times the i^{th} value occurs.

4. Compute the *z* value using

$$z_W = \frac{W - \mu_{W_n}}{\sigma_{W_n}}$$

For cases when *n* is less than 38, the significance level is found from a table of exact probabilities for the Wilcoxon test. When *n* is greater than or equal to 38, the significance of the test statistic is determined by computing the p-value using the standard normal distribution. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, no conclusion can be reached.

Sign Test

The sign test is popular because it is simple to compute. It assumes that the data follow the same distribution. The test is computed using the following steps.

1. Count the number of values strictly greater than *M*0. Call this value *X*.
2. Count the number of values strictly less than *M*0. Call this value *Y*.
3. Set $m = X + Y$.
4. Under the null hypothesis, *X* is distributed as a binomial random variable with a proportion of 0.5 and sample size of *m*.

The significance of *X* is calculated using binomial probabilities.

Bootstrap Test

The one-sample bootstrap procedure for testing whether the mean is equal to a specific value is given in Efron & Tibshirani (1993) pages 224-227. The bootstrap procedure is as follows.

1. Compute the mean of the sample. Call it $\overline{X}$.

2. Compute the t-value using the standard t-test. The formula for this computation is

$$t_X = \frac{\overline{X} - M0}{s_{\overline{X}}}$$

3. Draw a random, with-replacement sample of size *n* from the original *X* values. Call this sample $Y_1, Y_2, \cdots, Y_n$.

4. Compute the t-value of this bootstrap sample using the formula

$$t_Y = \frac{\overline{Y} - \overline{X}}{s_{\overline{Y}}}$$

5. For a two-tailed test, if $|t_Y| > |t_x|$ then add one to a counter variable *A*.

6. Repeat steps 3 – 5 *B* times. *B* may be anywhere from 100 to 10,000.

7. Compute the *p*-value of the bootstrap test as (*A* + 1) / (*B* + 1)

8. Steps 1 – 7 complete one simulation iteration. Repeat these steps *M* times, where *M* is the number of simulations. The power and significance level is equal to the percent of the time the *p*-value is less than the nominal alpha of the test.

Note that the bootstrap test is a time-consuming test to run, especially if you set *B* to a value larger than 100.

The Problem of Differing Standard Deviations

Care must be used when either the null or alternative distribution is not normal. In these cases, the standard deviation is usually not specified directly. For example, you might use a gamma distribution with a shape parameter of 1.5 and a mean of 4 as the null distribution and a gamma distribution with the same shape parameter and a mean of 5 as the alternative distribution. This allows you to compare the two means. However, note that although the shape parameters are constant, the standard deviations are not. Thus the null and alternatives not only have different means, but different standard deviations!

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data and Options tabs. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that will be of interest.

Solve For

Find (Solve For)

This option specifies whether you want to find *Power* or *N* from the simulation. Select *Power* when you want to estimate the power of a certain scenario. Select *N* when you want to determine the sample size needed to achieve a given power and alpha error level. Finding *N* is very computationally intensive, and so it may take a long time to complete.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies one or more values of the sample size, the number of subjects in the study. The paired design assumes that a pair of observations will be obtained from each subject. Thus there will be 2N observations simulated, resulting in N differences.

This value must be an integer greater than one. You may enter a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Test

Test Type

Specify which test statistic (t-test, Wilcoxon test, sign test, or bootstrap test) is to be simulated. Although the t-test is the most commonly used test statistic, it is based on assumptions that may not be viable in many situations. For your data, you may find that one of the other tests is more accurate (actual alpha = target alpha) and more precise (better power).

Note that the bootstrap test is computationally intensive, so it can be very slow to calculate.

Alternative Hypothesis

This option specifies the alternative hypothesis, H1. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always H0: Diff = Diff0.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **Difference <> Diff0**

  This is the most common selection. It yields a *two-tailed test*. Use this option when you are testing whether the mean is different from a specified value Diff0, but you do not want to specify beforehand whether it is smaller or larger. Most scientific journals require two-tailed tests.

- **Difference < Diff0**

  This option yields a *one-tailed test*. Use it when you want to test whether the true mean is less than Diff0.

- **Difference > Diff0**

  This option yields a *one-tailed test*. Use it when you want to test whether the true mean is greater than Diff0. Note that this option could be used for a **non-inferiority test**.

Simulations

Simulations

This option specifies the number of iterations, M, used in the simulation. As the number of iterations is increased, the accuracy and running time of the simulation will be increased also.

The precision of the simulated power estimates are calculated from the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the ‘Precision’ amount given in the table.

Simulation	Precision	Precision
Size	when	when
M	Power = 0.50	Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

Item A (and B) Distribution|H0

These options specify the distributions of the two items making up the pair under the null hypothesis, H0. The difference between the means of these two distributions is the difference that is tested, Diff0.

Usually, you will want Diff0 = 0. This zero difference is specified by entering *M0* for the mean parameter in each of the distributions and then entering an appropriate value for the M0 parameter below.

All of the distributions are parameterized so that the mean is entered first. For example, if you wanted to test whether the mean of a normal distributed variable is five, you could enter N(5, S) or N(M0, S) here.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them:

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and are not repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Item A (and B) Distribution|H1

These options specify the distributions of the two items making up the pair under the alternative hypothesis, H1. The difference between the means of these two distributions is the difference that is assumed to be the true value of the difference. That is, this is the difference at which the power is computed.

Usually, the mean difference is specified by entering *M1* for the mean parameter in the distribution expression for item A and *M0* for the mean parameter in the distribution expression for item B. The mean difference under H1 then becomes the value of M1 – M0.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them:

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for the *M0* in the four distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for the *M1* in the four distribution specifications given above. *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Effect Size – Distribution Parameters

R (Correlation of Items A & B)

Specify the value of the correlation between items (variates) A and B of the pair.

Since this is a correlation, it must be between -1 and 1. However, some distributions (such as the multinomial distribution) have a maximum possible correlation that is far less than one.

Typical values are between 0 and 0.4.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size, N, is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Bootstrap Iterations

Bootstrap Iterations

Specify the number of iterations used in the bootstrap hypothesis test. This value is only used if the bootstrap test is displayed on the reports. The running time of the procedure depends heavily on the number of iterations specified here.

Recommendations by authors of books discussing the bootstrap are from 100 to 10,000. If you enter a large (greater than 500) value, the simulation may take several hours to run.

Random Numbers

Random Number Pool Size

This is the size of the pool of random values from which the random samples will be drawn. Populations of at least 10,000 should be used. Also, the value should be about twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

Note that values over 50,000 may take a long time to permute to achieve the target means and correlation.

Correlation

Maximum Switches

This option specifies the maximum number of index switches that can be made while searching for a permutation of item B that yields a correlation within the specified range. A value near 5,000,000 may be necessary when the correlation is near one.

Correlation Tolerance

Specify the amount above and below the target correlation that will still let a particular index-permutation to be selected for the population. For example, if you have selected a correlation of 0.3 and you set this tolerance to 0.001, then only populations with a correlation between 0.299 and 0.301 will be used. The recommended is 0.001 or smaller. Valid values are between 0 and 0.999.

Example 1 – Power at Various Sample Sizes

Researchers are planning a pre-post experiment to test whether the difference in response to a certain drug is different from zero. The researchers will use a paired t-test with an alpha level of 0.05. They want to compare the power at sample sizes of 50, 100, and 150 when the shift in the means is 0.6 from pre-test to post-test. They assume that the data are normally distributed with a standard deviation of 2 and that the correlation between the pre-test and post-test values is 0.20. Since this is an exploratory analysis, they set the number of simulation iterations to 2000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

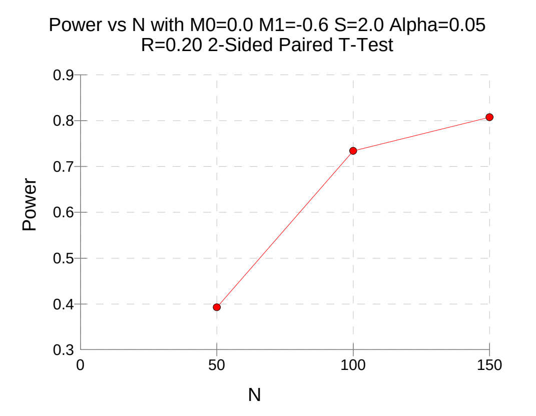
Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**50 100 150**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Item A Distribution|H0	**N(M0 S)**
Item B Distribution|H0	**N(M0 S)**
Item A Distribution|H1	**N(M0 S)**
Item B Distribution|H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**
S	**2**
R (Correlation of Items A & B)	**0.2**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Mean Difference = Diff0.    Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist'n: Normal(M0 S) - Normal(M0 S)
H1 Dist'n: Normal(M0 S) - Normal(M1 S)
Test Statistic: Paired T-Test

                          H0        H1      Corr    Target    Actual
Power         N        Diff0     Diff1         R     Alpha     Alpha      Beta        M0        M1         S
0.393        50          0.0      -0.6     0.200     0.050     0.055     0.608       0.0       0.6       2.0
(0.021)  [0.371        0.414]                                (0.010)    [0.045    0.064]

0.734       100          0.0      -0.6     0.200     0.050     0.050     0.266       0.0       0.6       2.0
(0.019)  [0.715        0.753]                                (0.010)    [0.040    0.060]

0.808       150          0.0      -0.6     0.200     0.050     0.058     0.193       0.0       0.6       2.0
(0.017)  [0.790        0.825]                                (0.010)    [0.048    0.068]

Notes:
Number of Monte Carlo Samples: 2000.   Simulation Run Time: 19.33 seconds.

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N is the size of the sample drawn from the population.
Diff0 is the paired-difference mean (A-B) assuming the null hypothesis, H0. This is the value being tested.
Diff1 is the paired-difference mean (A-B) assuming the alternative hypothesis, H1. This is the true value.
R is the correlation between the paired items.
Target Alpha is the probability of rejecting a true null hypothesis. It is set by the user.
Actual Alpha is the alpha level that was actually achieved by the experiment.
Beta is the probability of accepting a false null hypothesis.
Second Row: (Power Inc.) [95% LCL and UCL Power]    (Alpha Inc.) [95% LCL and UCL Alpha]

Summary Statements
A sample size of 50 achieves 39% power to detect a difference of -0.6 between the null
hypothesis mean difference of 0.0 and the actual mean difference of -0.6 at the 0.050
significance level (alpha) using a two-sided Paired T-Test. These results are based on 2000
Monte Carlo samples from the null distribution: Normal(M0 S) - Normal(M0 S) and the alternative
distribution: Normal(M0 S) - Normal(M1 S).

Plots Section
```



This report shows the estimated power for each scenario. The first row shows the parameter settings and the estimated power and significance level (Actual Alpha).

The second row shows two 95% confidence intervals in brackets: the first for the power and the second for the significance level. Half the width of each confidence interval is given in parentheses as a fundamental measure of the accuracy of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Example 2 – Finding the Sample Size

Continuing with Example 1, the researchers want to determine how large a sample is needed to obtain a power of 0.90?

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.90**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Item A Distribution|H0	**N(M0 S)**
Item B Distribution|H0	**N(M0 S)**
Item A Distribution|H1	**N(M0 S)**
Item B Distribution|H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**
S	**2**
R (Correlation of Items A & B)	**0.2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results of Search for N

```
                        H0        H1      Corr    Target    Actual
Power        N       Diff0     Diff1         R     Alpha     Alpha      Beta        M0        M1         S
0.885      193         0.0      -0.6     0.200     0.050     0.057     0.115       0.0       0.6       2.0
(0.016)  [0.869     0.901]                                (0.010)    [0.046    0.067]
Notes:
Number of Monte Carlo Samples: 2000.   Simulation Run Time: 95.53 seconds.
```

The required sample size of 193 achieved a power of 0.885. The power of 0.885 is less than the target value of 0.900 because the sample size search algorithm re-simulates the power for the final sample size. Thus it is possible for the search algorithm to converge to a sample size which exhibits the desired power, but then on a succeeding simulation to achieve a power that is slightly less than the target. To achieve more accuracy, a reasonable strategy would be to run simulations to obtain the powers using N's from 190 to 200 using a simulation size of 5000.

Example 3 – Comparative Results

Continuing with Example 2, the researchers want to study the characteristics of alternative test statistics.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**50 100 150 200**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**2000**
Item A Distribution|H0	**N(M0 S)**
Item B Distribution|H0	**N(M0 S)**
Item A Distribution|H1	**N(M0 S)**
Item B Distribution |H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**

Data Tab (continued)

S **2**
R (Correlation of Items A & B) **0.2**

Reports Tab

Show Comparative Reports .... **Checked**
Show Comparative Plots......................... **Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist'n: Normal(M0 S) - Normal(M0 S)
H1 Dist'n: Normal(M0 S) - Normal(M1 S)

              H0        H1
            Diff      Diff      Corr    Target    T-Test    Wilcxn      Sign
N         (Diff0)   (Diff1)      (R)     Alpha     Power     Power     Power        M0        M1         S
50            0.0      -0.6    0.200     0.050     0.367     0.356     0.206       0.0       0.6       2.0
100           0.0      -0.6    0.200     0.050     0.661     0.664     0.467       0.0       0.6       2.0
150           0.0      -0.6    0.200     0.050     0.755     0.740     0.532       0.0       0.6       2.0
200           0.0      -0.6    0.200     0.050     0.960     0.960     0.849       0.0       0.6       2.0
Number of Monte Carlo Iterations: 2000.   Simulation Run Time: 36.70 seconds.

Alpha Comparison for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1<>Diff0
H0 Dist'n: Normal(M0 S) - Normal(M0 S)
H1 Dist'n: Normal(M0 S) - Normal(M1 S)

              H0        H1
            Diff      Diff      Corr    Target    T-Test    Wilcxn      Sign
N         (Diff0)   (Diff1)      (R)     Alpha     Alpha     Alpha     Alpha        M0        M1         S
50            0.0      -0.6    0.200     0.050     0.062     0.060     0.041       0.0       0.6       2.0
100           0.0      -0.6    0.200     0.050     0.046     0.045     0.040       0.0       0.6       2.0
150           0.0      -0.6    0.200     0.050     0.045     0.049     0.047       0.0       0.6       2.0
200           0.0      -0.6    0.200     0.050     0.041     0.039     0.041       0.0       0.6       2.0
Number of Monte Carlo Iterations: 2000.   Simulation Run Time: 36.70 seconds.
```

These results show that for paired data, the t-test and Wilcoxon test have very similar power and alpha values. The sign test is less accurate and less powerful.

Example 4 – Validation

We will validate this procedure by comparing its results to those of the regular one-sample t-test, a procedure that has already by validated. For this run, we will use the settings of Example 1: M0 = 0, M1 = 0.6, alpha = 0.05, N = 50, R = 0.2, and S = 2.

Note that to run this example using the regular one-sample t-test procedure, the variance will have to be altered to account for the correlation of 0.20. The adjusted standard deviation is equal to S times the square root of 2(1 – R), which, in this case, is 2.530. Running this through the regular One Mean procedure yields a power of 0.376.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**50**
Test Type	**T-Test**
Alternative Hypothesis	**Diff<>Diff0**
Simulations	**10000**
Item A Distribution|H0	**N(M0 S)**
Item B Distribution|H0	**N(M0 S)**
Item A Distribution|H1	**N(M0 S)**
Item B Distribution|H1	**N(M1 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.6**
S	**2**
R (Correlation of Items A & B)	**0.2**
Options Tab	
Random Number Pool Size	**50000 (Increase to 5 times Simulations)**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                        H0        H1      Corr   Target    Actual
Power          N     Diff0     Diff1         R    Alpha     Alpha      Beta        M0        M1         S
0.373         50       0.0      -0.6     0.200    0.050     0.049     0.627       0.0       0.6       2.0
(0.009)  [0.363     0.382]                                (0.004)    [0.045    0.053]

Notes:
Number of Monte Carlo Samples: 10000.   Simulation Run Time: 30.97 seconds.
```

The power matches the exact value of 0.376 quite well. We re-ran the procedure several times and obtained power values from 0.370 to 0.396.

Example 5 – Non-Inferiority Test

A non-inferiority test is appropriate when you want to show that a new treatment is no worse than the standard. For example, suppose that a standard diagnostic test has an average score of 70. Unfortunately, this diagnostic test is expensive. A promising new diagnostic test must be compared to the standard. Researchers want to show that it is no worse than the standard.

Because of many benefits from the new test, clinicians are willing to adopt it even if it is slightly less accurate than the current test. How much less can the score of the new treatment be and still be adopted? Should it be adopted if the difference is -1? -2? -5? -10? There is an amount below 0 at which the difference between the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, the *margin of equivalence* is set to -5.

The developers decided to use a paired t-test. They must design an experiment to test the hypothesis that the average difference between the two tests is greater than -5. The statistical hypothesis to be tested is

$$H_0 : A - B \leq -5 \text{ versus } H_1 : A - B > -5$$

where A represents the mean of the new test and B represents the mean of the standard test. Notice that when the null hypothesis is rejected, the conclusion is that the average difference is greater than -5.

Past experience has shown that the standard deviation is 5.0 and the correlation is 0.2. Following proper procedure, the researchers decide to use a significance level of 0.025 for this one-sided test to keep it comparable to the usual value of 0.05 for a two-sided test. They decide to look at the power for sample sizes of 5, 10, 15, 20, and 25 subjects. They decide to compute the power for the case when the two tests are actually equal.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Inequality Tests**, then **Specify using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N (Sample Size)	**5 10 15 20 25**
Test Type	**T-Test**
Alternative Hypothesis	**Diff>Diff0**
Simulations	**2000**
Item A Distribution|H0	**N(M0 S)**
Item B Distribution|H0	**N(M1 S)**
Item A Distribution|H1	**N(M0 S)**
Item B Distribution|H1	**N(M0 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**5**
S	**5**
R (Correlation of Items A & B)	**0.2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Mean Difference = Diff0.   Hypotheses: H0: Diff1=Diff0; H1: Diff1>Diff0
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)
Test Statistic: Paired T-Test

                    H0       H1     Corr   Target   Actual
Power       N     Diff0    Diff1       R    Alpha    Alpha     Beta       M0       M1        S
0.308       5      -5.0      0.0   0.200    0.025    0.023    0.692      0.0      5.0      5.0
(0.020) [0.288    0.328]                           (0.007)   [0.016   0.030]

0.617      10      -5.0      0.0   0.200    0.025    0.024    0.383      0.0      5.0      5.0
(0.021) [0.596    0.638]                           (0.007)   [0.017   0.030]

0.816      15      -5.0      0.0   0.200    0.025    0.027    0.184      0.0      5.0      5.0
(0.017) [0.799    0.833]                           (0.007)   [0.019   0.034]

0.916      20      -5.0      0.0   0.200    0.025    0.022    0.085      0.0      5.0      5.0
(0.012) [0.903    0.928]                           (0.006)   [0.016   0.028]

0.968      25      -5.0      0.0   0.200    0.025    0.025    0.032      0.0      5.0      5.0
(0.008) [0.960    0.976]                           (0.007)   [0.018   0.031]

Notes:
Number of Monte Carlo Samples: 2000.   Simulation Run Time: 13.34 seconds.
```

We see that a power of 0.8 is achieved at about 15 subjects, while a power of 0.9 requires about 20 subjects.

Chapter 495

Equivalence Tests for Paired Means (Simulation)

Introduction

This procedure allows you to study the power and sample size of tests of equivalence of means of two correlated variables. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. The paired t-test is commonly used in this situation. Other tests have been developed for the case when the data are not normally distributed. These additional tests include the Wilcoxon signed-ranks test, the sign test, and the computer-intensive bootstrap test.

Paired data may occur because two measurements are made on the same subject or because measurements are made on two subjects that have been matched according to other, often demographic, variables. Hypothesis tests on paired data can be analyzed by considering the differences between the paired items. The distribution of differences is usually symmetric. In fact, the distribution must be symmetric if the individual distributions of the two items are identical. Hence, the paired t-test and the Wilcoxon signed-rank test are appropriate for paired data even when the distributions of the individual items are not normal.

The details of the power analysis of the paired t-test using analytic techniques are presented in another ***PASS*** chapter and they won't be duplicated here. This chapter will only consider power analysis using computer simulation.

Technical Details

Computer simulation allows us to estimate the power and significance-level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are as follows.

1. Specify the test procedure and the test statistic. This includes the significance level, sample size, and underlying data distributions.

2. Generate a random sample $X_1, X_2, \ldots, X_n$ from the distribution specified by the alternative hypothesis. Calculate the test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. These samples are used to calculate the power of the test. In the case of paired data, the individual values are simulated as the difference between two other random variables. These samples are constructed so that they exhibit a certain amount of correlation.

3. Generate a random sample $Y_1, Y_2, \ldots, Y_n$ from the distribution specified by the null hypothesis. Calculate the test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. These samples are used to calculate the significance-level of the test. In the case of paired data, the individual values are simulated as the difference between two other random variables. These samples are constructed so that they exhibit a certain amount of correlation.

4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulation samples in step 2 that lead to rejection. The significance-level is the proportion of simulated samples in step 3 that lead to rejection.

Simulating Paired Distributions

Paired data occur when two observations are correlated. Examples of paired designs are pre – post designs, cross-over designs, and matched pair designs.

In order to simulate paired data, the simulation should mimic the actual data generation process as closely as possible. Since paired data are analyzed by creating the individual difference between each pair, the simulation should also create data as the difference between two variates. Paired data exhibit a correlation between the two variates. As this correlation between the variates increases, the variance of the difference decreases. Thus it is important not only to specify the distributions of the two variates that will be differenced, but to also specify their correlation.

Obtaining paired samples from arbitrary distributions with a set correlation is difficult because the joint, bivariate distribution must be specified and simulated. Rather than specify the bivariate distribution, ***PASS*** requires the specification of the two marginal distributions and the correlation between them.

Monte Carlo samples with given marginal distributions and correlation are generated using the method suggested by Gentle (1998). The method begins by generating a large population of random numbers from the two distributions. Each of these populations is evaluated to determine if their means are within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean.

The next step is to obtain the target correlation. This is accomplished by permuting one of the populations until they have the desired correlation.

The above steps provide a large pool of random numbers that exhibit the desired characteristics. This pool is then sampled at random using the uniform distribution to obtain the random numbers used in the simulation.

This algorithm may be stated as follows.

1. Draw individual samples of size M from the two distributions where M is a large number, usually over 10,000. Adjust these samples so that they have the specified mean and standard deviation. Label these samples A and B. Create an index of the values of A and B according to the order in which they are generated. Thus, the first value of A and the first value of B are indexed as one, the second values of A and B are indexed as two, and so on up to the final set which is indexed as M.
2. Compute the correlation between the two generated variates.
3. If the computed correlation is within a small tolerance (usually less than 0.001) of the specified correlation, go to step 7.
4. Select two indices (I and J) at random using uniform random numbers.
5. Determine what will happen to the correlation if B_I is swapped with B_J. If the swap will result in a correlation that is closer to the target value, swap the indices and proceed to step 6. Otherwise, go to step 4.
6. If the computed correlation is within the desired tolerance of the target correlation, go to step 7. Otherwise, go to step 4.
7. End with a population with the required marginal distributions and correlation.

Now, to complete the simulation, random samples of the designated size are drawn from this population.

Simulating Data for an Equivalence Test

Simulating equivalence data is more complex than simulating data for a regular two-sided test. An equivalence test essentially reverses the roles of the null and alternative hypothesis. In so doing, the null hypothesis becomes

$$H0\colon (\mu_1 - \mu_2) \le -D \ or \ (\mu_1 - \mu_2) \ge D$$

where *D* is the margin of equivalence. Thus the null hypothesis is made up of two simple hypotheses:

$$H0_1\colon (\mu_1 - \mu_2) \le -D$$

$$H0_2\colon (\mu_1 - \mu_2) \ge D$$

The additional complexity comes in deciding which of the two simple null hypotheses are used to simulate data for the null hypothesis situation. The choice becomes more problematic when asymmetric equivalence limits are chosen. In that case, you may want to try simulating using each simple null hypothesis in turn.

To generate data for the null hypotheses, you generate data for each group. The difference in the means of these two groups will become one of the equivalence limits. The other equivalence limit will be determined by symmetry and will always have a sign that is the negative of the first equivalence limit.

Test Statistics

This section describes the test statistics that are available in this procedure. Note that these test statistics are computed on the differences. Thus, when the equation refers to an X value, this X value is assumed to be a difference between two individual variates.

One-Sample t-Test

The one-sample t-test assumes that the data are a simple random sample from a population of normally-distributed values that all have the same mean and variance. This assumption implies that the data are continuous and their distribution is symmetric. The calculation of the t-test proceeds as follow

$$t_{n-1} = \frac{\overline{X} - M0}{s_{\overline{X}}}$$

where

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

$$s_{\overline{X}} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}},$$

and $M0$ is the value of the difference hypothesized by the null hypothesis.

The significance of the test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the hypothesis is rejected. Otherwise, no conclusion can be reached.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a popular, nonparametric substitute for the t-test. It assumes that the data follow a symmetric distribution. The test is computed using the following steps.

1. Subtract the hypothesized mean, $D0$, from each data value. Rank the values according to their absolute values.
2. Compute the sum of the positive ranks *Sp* and the sum of the negative ranks *Sn*. The test statistic, *W*, is the minimum of *Sp* and *Sn*.
3. Compute the mean and standard deviation of *W* using the formulas

$$\mu_{W_n} = \frac{n(n+1)}{4} \text{ and } \sigma_{W_n} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}$$

where t_i represents the number of times the i^{th} value occurs.

4. Compute the *z* value using

$$z_W = \frac{W - \mu_{W_n}}{\sigma_{W_n}}$$

For cases when *n* is less than 38, the significance level is found from a table of exact probabilities for the Wilcoxon test. When *n* is greater than or equal to 38, the significance of the test statistic is determined by computing the p-value using the standard normal distribution. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, no conclusion can be reached.

Sign Test

The sign test is popular because it is simple to compute. It assumes that the data follow the same distribution. The test is computed using the following steps.

1. Count the number of values strictly greater than *M*0. Call this value *X*.
2. Count the number of values strictly less than *M*0. Call this value *Y*.
3. Set $m = X + Y$.
4. Under the null hypothesis, *X* is distributed as a binomial random variable with a proportion of 0.5 and sample size of *m*.

The significance of *X* is calculated using binomial probabilities.

Bootstrap Test

The one-sample bootstrap procedure for testing whether the mean is equal to a specific value is given in Efron & Tibshirani (1993) pages 224-227. The bootstrap procedure is as follows.

1. Compute the mean of the sample. Call it $\overline{X}$.
2. Compute the t-value using the standard t-test. The formula for this computation is

$$t_X = \frac{\overline{X} - M0}{s_{\overline{X}}}$$

3. Draw a random, with-replacement sample of size *n* from the original *X* values. Call this sample $Y_1, Y_2, \cdots, Y_n$.
4. Compute the t-value of this bootstrap sample using the formula

$$t_Y = \frac{\overline{Y} - \overline{X}}{s_{\overline{Y}}}$$

5. For a two-tailed test, if $|t_Y| > |t_x|$ then add one to a counter variable *A*.
6. Repeat steps 3 – 5 *B* times. *B* may be anywhere from 100 to 10,000.
7. Compute the *p*-value of the bootstrap test as (*A* + 1) / (*B* + 1)
8. Steps 1 – 7 complete one simulation iteration. Repeat these steps *M* times, where *M* is the number of simulations. The power and significance level is equal to the percent of the time the *p*-value is less than the nominal alpha of the test.

Note that the bootstrap test is a time-consuming test to run, especially if you set *B* to a value larger than 100.

The Problem of Differing Standard Deviations

Care must be used when either the null or alternative distribution is not normal. In these cases, the standard deviation is usually not specified directly. For example, you might use a gamma distribution with a shape parameter of 1.5 and a mean of 4 as the null distribution and a gamma distribution with the same shape parameter and a mean of 5 as the alternative distribution. This allows you to compare the two means. However, note that although the shape parameters are constant, the standard deviations are not. Thus the null and alternatives not only have different means, but different standard deviations!

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data and Options tabs. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies whether you want to find *Power* or *N* from the simulation. Select *Power* when you want to estimate the power of a certain scenario. Select *N* when you want to determine the sample size needed to achieve a given power and alpha error level. Finding *N* is very computationally intensive, and so it may take a long time to complete.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

This option specifies one or more values of the sample size, the number of subjects in the study. The paired design assumes that a pair of observations will be obtained from each subject. Thus there will be 2N observations simulated, resulting in N differences.

This value must be an integer greater than one. You may enter a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Test

Test Type

Specify which test statistic (t-test, Wilcoxon test, sign test, or bootstrap test) is to be simulated. Although the t-test is the most commonly used test statistic, it is based on assumptions that may not be viable in many situations. For your data, you may find that one of the other tests is more accurate (actual alpha = target alpha) and more precise (better power).

Note that the bootstrap test is computationally intensive, so it can be very slow to calculate.

Equivalence Limit

Equivalence limits are defined as the positive and negative limits around zero that define a zone of equivalence. This zone of equivalence is a set of difference values that define a region in which the two means are 'close enough' so that they are considered to be the same for practical purposes.

Rather than define these limits explicitly, they are set implicitly. This is done as follows. One limit is found by subtracting the Item B mean | H0 from the Item A mean | H0. If the limits are symmetric, the other limit is this difference times -1. To obtain symmetric limits, enter 'Symmetric' here.

If asymmetric limits are desired, a numerical value is specified here. It will be given the sign (+ or -) that is opposite the difference in the means discussed above.

For example, if the mean of A under H0 is 5, the mean of B under H0 is 4, and 'Symmetric' is entered here, the equivalence limits will be 5 - 4 = 1 and -1. However, if the value '1.25' is entered here, the equivalence limits are 1 and -1.25.

If you do not have a specific value in mind for the equivalence limit, a common value for an equivalence limit is 20% or 25% of the Item A (reference) mean.

Simulations

Simulations

This option specifies the number of iterations, M, used in the simulation. As the number of iterations is increased, the accuracy and running time of the simulation will be increased also.

The precision of the simulated power estimates are calculated from the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the ‘Precision’ amount given in the table.

Simulation Size M	Precision when Power = 0.50	Precision when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

Item A (and B) Distribution|H0

These options specify the distributions of the two items making up the pair under the null hypothesis, H0. The difference between the means of these two distributions is the difference that is tested, Diff0.

Usually, you will want Diff0 = 0. This zero difference is specified by entering *M0* for the mean parameter in each of the distributions and then entering an appropriate value for the M0 parameter below.

All of the distributions are parameterized so that the mean is entered first. For example, if you wanted to test whether the mean of a normal distributed variable is five, you could enter N(5, S) or N(M0, S) here.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them:

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)

Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and are not repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Item A (and B) Distribution|H1

These options specify the distributions of the two items making up the pair under the alternative hypothesis, H1. The difference between the means of these two distributions is the difference that is assumed to be the true value of the difference. That is, this is the difference at which the power is computed.

Usually, the mean difference is specified by entering *M1* for the mean parameter in the distribution expression for item A and *M0* for the mean parameter in the distribution expression for item B. The mean difference under H1 then becomes the value of M1 – M0.

The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them:

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for the *M0* in the four distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for the *M1* in the four distribution specifications given above. *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

R (Correlation of Items A & B)

Specify the value of the correlation between items (variates) A and B of the pair.

Since this is a correlation, it must be between -1 and 1. However, some distributions (such as the multinomial distribution) have a maximum possible correlation that is far less than one.

Typical values are between 0 and 0.4.

Parameter Values (S, A, B)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size, N, is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Bootstrap Iterations

Bootstrap Iterations

Specify the number of iterations used in the bootstrap hypothesis test. This value is only used if the bootstrap test is displayed on the reports. The running time of the procedure depends heavily on the number of iterations specified here.

Recommendations by authors of books discussing the bootstrap are from 100 to 10,000. If you enter a large (greater than 500) value, the simulation may take several hours to run.

Random Numbers

Random Number Pool Size

This is the size of the pool of random values from which the random samples will be drawn. Populations of at least 10,000 should be used. Also, the value should be about twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

Note that values over 50,000 may take a long time to permute to achieve the target means and correlation.

Correlation

Maximum Switches

This option specifies the maximum number of index switches that can be made while searching for a permutation of item B that yields a correlation within the specified range. A value near 5,000,000 may be necessary when the correlation is near one.

Correlation Tolerance

Specify the amount above and below the target correlation that will still let a particular permutation to be selected for the population. For example, if you have selected a correlation of 0.3 and you set this tolerance to 0.001, then only populations with a correlation between 0.299 and 0.301 will be used. The recommended is 0.001 or smaller. Valid values are between 0 and 0.999.

Example 1 – Power at Various Sample Sizes

Researchers are planning an experiment to determine if the response to a new drug is equivalent to the response to the standard drug. The average response level to the standard drug is 63 with a standard deviation of 5. The researchers decide that if the average response level to the new drug is between 60 and 66, they will consider it to be equivalent to the standard drug.

The researchers decide to use a paired design so that each subject can serve as their own control. The response level for the standard drug will be measured for each subject. Then, followed by an appropriate wash-out period of two days, the response level to the new drug will be measured. From previous studies, they know that the correlation between the two response levels will be between 0.1 and 0.20.

The researchers will analyze the data using an equivalence test based on the paired t-test with an alpha level of 0.05. They want to compare the power at sample sizes of 10, 30, 50, and 70. They assume that the data are normally distributed and that the true difference between the response level of the two drugs is zero. Since this is an exploratory analysis, they set the number of simulation iterations to 2000.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Equivalence Tests using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**10 30 50 70**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Item A (Reference) Dist'n|H0	**N(M0 S)**
Item B (Treatment) Dist'n|H0	**N(M1 S)**
Item A (Reference) Dist'n|H1	**N(M0 S)**
Item B (Treatment) Dist'n|H1	**N(M0 S)**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**
R (Correlation of Items A & B)	**0.1 0.2**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

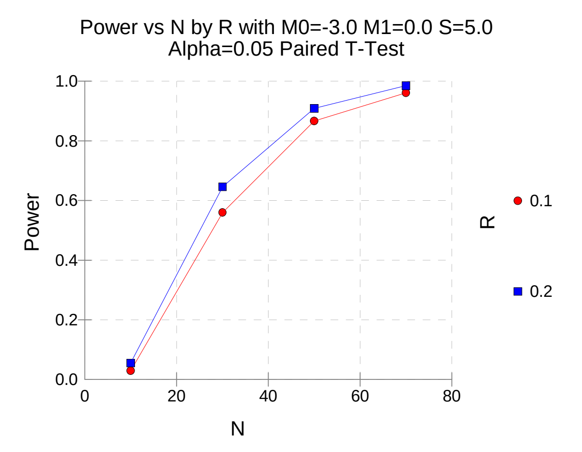
Numeric Results

```
Numeric Results for Testing Mean Equivalence.    Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)
Test Statistic: Paired T-Test

                            Lower    Upper
                     H1    Equiv.   Equiv.     Corr   Target   Actual
Power        N    Diff1    Limit    Limit        R    Alpha    Alpha       M0        M1         S
0.030       10      0.0     -3.0      3.0    0.100    0.050    0.009     63.0      66.0       5.0
(0.007)  [0.022  0.037]                                      (0.004)   [0.005    0.013]

0.055       10      0.0     -3.0      3.0    0.200    0.050    0.019     63.0      66.0       5.0
(0.010)  [0.045  0.065]                                      (0.006)   [0.013    0.025]

0.560       30      0.0     -3.0      3.0    0.100    0.050    0.050     63.0      66.0       5.0
(0.022)  [0.538  0.581]                                      (0.010)   [0.040    0.060]

Population Size: 10000. Number of Monte Carlo Samples: 2000.   Simulation Run Time: 30.02 seconds.
```

Report Definitions
Power is the probability of rejecting a false null hypothesis.
N is the size of the sample drawn from the population.
Diff1 is the paired-difference mean (A-B) assuming the alternative hypothesis, H1. This is the true value.
Lower Equiv Limit is the lower limit on a difference (A-B) that is considered as equivalent.
Upper Equiv Limit is the upper limit on a difference (A-B) that is considered as equivalent.
Diff0 is the paired-difference mean (A-B) assuming the null hypothesis, H0. This is one of the equivalence limits.
R is the correlation between the paired items.
Target Alpha is the probability of rejecting a true null hypothesis. It is set by the user.
Actual Alpha is the alpha level that was actually achieved by the experiment.
Beta is the probability of accepting a false null hypothesis.
Second Row: (Power Inc.) [95% LCL and UCL Power]   (Alpha Inc.) [95% LCL and UCL Alpha]

Summary Statements
A sample size of 10 pairs with a correlation of 0.100 achieves 3% power to detect equivalence when the margin of equivalence is from -3.0 to 3.0 and the actual mean difference is 0.0. The significance level (alpha) is 0.050 using two one-sided Paired T-Tests. These results are based on 2000 Monte Carlo samples from the null distribution: Normal(M0 S) - Normal(M1 S) and the alternative distribution: Normal(M0 S) - Normal(M0 S).

Chart Section



This report shows the estimated power for each scenario. The first row shows the parameter settings and the estimated power and significance level (Actual Alpha).

The second row shows two 95% confidence intervals in brackets: the first for the power and the second for the significance level. Half the width of each confidence interval is given in parentheses as a fundamental measure of the accuracy of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

We see that a sample size of about 50 is needed to obtain a reasonable power level.

Example 2 – Finding the Sample Size

Continuing with Example 1, the researchers want to determine how large a sample is needed to obtain a power of 0.90? They decide to use a correlation of 0.10, since that will result in a larger, more conservative, sample size.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Equivalence Tests using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.9**
Alpha	**0.05**
N (Sample Size)	*Ignored since this is the Find setting*
Simulations	**2000**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Item A (Reference) Dist'n|H0	**N(M0 S)**
Item B (Treatment) Dist'n|H0	**N(M1 S)**
Item A (Reference) Dist'n|H1	**N(M0 S)**
Item B (Treatment) Dist'n|H1	**N(M0 S)**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**
R (Correlation of Items A & B)	**0.1 0.2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Mean Equivalence.    Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)
Test Statistic: Paired T-Test

                                Lower     Upper
                        H1     Equiv.    Equiv.      Corr   Target    Actual
Power          N      Diff1     Limit     Limit         R    Alpha     Alpha        M0         M1        S
0.899         54        0.0      -3.0       3.0     0.100    0.050     0.044      63.0       66.0      5.0
(0.013)   [0.885     0.912]                                          (0.009)    [0.035     0.052]
```

The required sample size was 54 which achieved a power of 0.899.

The power of 0.899 is slightly less than the target value of 0.900 because the sample size search algorithm re-simulates the power for the final sample size. Thus it is possible for the search algorithm to converge to a sample size which exhibits the desired power, but then on the second simulation, achieves a power that is slightly less than the target. To obtain more accuracy, a reasonable strategy would be to run simulations to obtain the powers using N's from 50 to 60 using a simulation size of 5000.

Example 3 – Comparing Test Statistics

Continuing with Example 2, the researchers want to study the characteristics of alternative test statistics.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Equivalence Tests using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

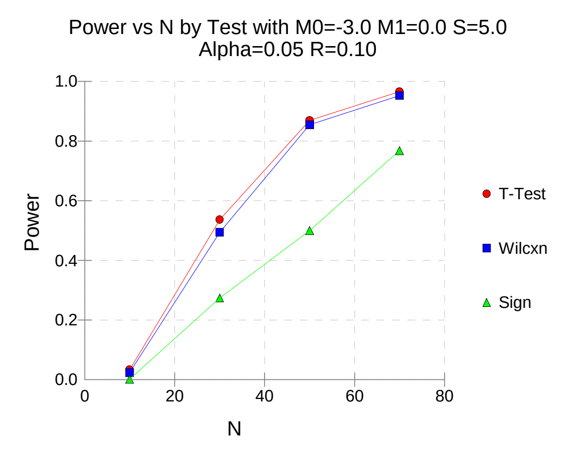
Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**10 30 50 70**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Item A (Reference) Dist'n|H0	**N(M0 S)**
Item B (Treatment) Dist'n|H0	**N(M1 S)**
Item A (Reference) Dist'n|H1	**N(M0 S)**
Item B (Treatment) Dist'n|H1	**N(M0 S)**
M0 (Mean|H0)	**63**
M1 (Mean|H1)	**66**
S	**5**
R (Correlation of Items A & B)	**0.1**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)

            H1     Lower    Upper
          Diff    Equiv.   Equiv.     Corr   Target   T-Test   Wilcxn     Sign
N       (Diff1)    Limit    Limit      (R)    Alpha    Power    Power    Power       M0        M1        S
10          0.0     -3.0      3.0    0.100    0.050    0.034    0.024    0.002     63.0      66.0      5.0
30          0.0     -3.0      3.0    0.100    0.050    0.537    0.494    0.275     63.0      66.0      5.0
50          0.0     -3.0      3.0    0.100    0.050    0.870    0.855    0.500     63.0      66.0      5.0
70          0.0     -3.0      3.0    0.100    0.050    0.966    0.953    0.768     63.0      66.0      5.0

Alpha Comparison for Testing Equivalence.   Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)

            H1     Lower    Upper
          Diff    Equiv.   Equiv.     Corr   Target   T-Test   Wilcxn     Sign
N       (Diff1)    Limit    Limit      (R)    Alpha    Alpha    Alpha    Alpha       M0        M1        S
10          0.0     -3.0      3.0    0.100    0.050    0.010    0.009    0.001     63.0      66.0      5.0
30          0.0     -3.0      3.0    0.100    0.050    0.057    0.056    0.052     63.0      66.0      5.0
50          0.0     -3.0      3.0    0.100    0.050    0.052    0.049    0.024     63.0      66.0      5.0
70          0.0     -3.0      3.0    0.100    0.050    0.044    0.045    0.041     63.0      66.0      5.0
```



These results show that for paired data, the t-test and Wilcoxon test have very similar power and alpha values. The sign test is less accurate and less powerful.

Example 4 – Validation using Chow et al.

We will validate this procedure by comparing its results to those of Chow et al. (2003) page 55 in which the parameter values are: M0 = 0, M1 = 0.05, alpha = 0.05, N = 35, R = 0.0, and S = 0.070711. For these parameters, the power is given as 0.800.

Note that they give the standard deviation of the differences as 0.1. Since the correlation is 0.0, the standard deviation of the individual data values is given by 0.1/Sqrt(2) = 0.070711.

In order to understand the accuracy of the simulation, we will re-run the analysis five times.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Paired Means (Simulation)** procedure window by clicking on **Means**, then **Two Means**, then **Correlated (Paired)**, then **Equivalence Tests using Differences (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Sample Size)	**35 35 35 35 35**
Test Type	**T-Test**
Equivalence Limit	**Symmetric**
Simulations	**2000**
Item A (Reference) Dist'n|H0	**N(M0 S)**
Item B (Treatment) Dist'n|H0	**N(M1 S)**
Item A (Reference) Dist'n|H1	**N(M0 S)**
Item B (Treatment) Dist'n|H1	**N(M0 S)**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**0.05**
S	**0.070711**
R (Correlation of Items A & B)	**0.0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Mean Equivalence.    Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)
Test Statistic: Paired T-Test

                               Lower    Upper
                        H1    Equiv.   Equiv.     Corr   Target   Actual
Power          N     Diff1     Limit    Limit        R    Alpha    Alpha       M0        M1        S
0.813         35       0.0      -0.1      0.1    0.000    0.050    0.050      0.0       0.1      0.1
(0.017)  [0.795     0.830]                                       (0.010)   [0.040    0.059]

0.813         35       0.0      -0.1      0.1    0.000    0.050    0.045      0.0       0.1      0.1
(0.017)  [0.796     0.830]                                       (0.009)   [0.036    0.054]

0.803         35       0.0      -0.1      0.1    0.000    0.050    0.051      0.0       0.1      0.1
(0.017)  [0.785     0.820]                                       (0.010)   [0.041    0.061]

0.799         35       0.0      -0.1      0.1    0.000    0.050    0.045      0.0       0.1      0.1
(0.018)  [0.781     0.816]                                       (0.009)   [0.035    0.054]

0.826         35       0.0      -0.1      0.1    0.000    0.050    0.051      0.0       0.1      0.1
(0.017)  [0.809     0.843]                                       (0.010)   [0.041    0.060]

Notes:
Population Size: 10000. Number of Monte Carlo Samples: 2000.   Simulation Run Time: 16.80 seconds.
```

The powers match the analytic value of 0.800 quite well. Note how informative the confidence intervals are.

Chapter 496

Confidence Intervals for Paired Means

Introduction

This routine calculates the sample size necessary to achieve a specified distance from the paired sample mean difference to the confidence limit(s) at a stated confidence level for a confidence interval about the mean difference when the underlying data distribution is normal.

Caution: This procedure assumes that the standard deviation of the future sample will be the same as the standard deviation that is specified. If the standard deviation to be used in the procedure is estimated from a previous paired sample or represents the population standard deviation, the Confidence Intervals for Paired Means with Tolerance Probability procedure should be considered. That procedure controls the probability that the distance from the mean paired difference to the confidence limits will be less than or equal to the value specified.

Technical Details

For a paired sample mean difference from a normal distribution with known variance, a two-sided, 100(1 – α)% confidence interval is calculated by

$$\overline{X}_{Diff} \pm \frac{z_{1-\alpha/2}\sigma_{diff}}{\sqrt{n}}$$

where $\overline{X}_{Diff}$ is the mean of the paired differences of the sample, and σ_{diff} is the known standard deviation of paired sample differences.

A one-sided 100(1 – α)% upper confidence limit is calculated by

$$\overline{X}_{Diff} + \frac{z_{1-\alpha}\sigma_{diff}}{\sqrt{n}}$$

Similarly, the one-sided 100(1 – α)% lower confidence limit is

$$\overline{X}_{Diff} - \frac{z_{1-\alpha}\sigma_{diff}}{\sqrt{n}}$$

For a paired sample mean difference from a normal distribution with unknown variance, a two-sided, 100(1 – α)% confidence interval is calculated by

$$\overline{X}_{Diff} \pm \frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{diff}}{\sqrt{n}}$$

where $\overline{X}_{Diff}$ is the mean of the paired differences of the sample, and $\hat{\sigma}_{diff}$ is the estimated standard deviation of paired sample differences.

A one-sided 100(1 – α)% upper confidence limit is calculated by

$$\overline{X}_{Diff} + \frac{t_{1-\alpha,n-1}\hat{\sigma}_{diff}}{\sqrt{n}}$$

Similarly, the one-sided 100(1 – α)% lower confidence limit is

$$\overline{X}_{Diff} - \frac{t_{1-\alpha,n-1}\hat{\sigma}_{diff}}{\sqrt{n}}$$

Each confidence interval is calculated using an estimate of the mean difference plus and/or minus a quantity that represents the distance from the mean difference to the edge of the interval. For two-sided confidence intervals, this distance is sometimes called the precision, margin of error, or half-width. We will label this distance, *D*.

The basic equation for determining sample size when D has been specified is

$$D = \frac{z_{1-\alpha/2}\sigma_{diff}}{\sqrt{n}}$$

when the standard deviation is known, and

$$D = \frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{diff}}{\sqrt{n}}$$

when the standard deviation is unknown. These equations can be solved for any of the unknown quantities in terms of the others. The value $\alpha/2$ is replaced by α when a one-sided interval is used.

Finite Population Size

The above calculations assume that samples are being drawn from a large (infinite) population. When the population is of finite size (*N*), an adjustment must be made. The adjustment reduces the standard deviation as follows:

$$\sigma_{finite} = \sigma\sqrt{\left(1 - \frac{n}{N}\right)}$$

This new standard deviation replaces the regular standard deviation in the above formulas.

Confidence Level

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is $1 - \alpha$.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters.

Confidence

Confidence Level

The confidence level, $1 - \alpha$, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is $1 - \alpha$.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90, 0.95* or *0.90 to 0.99 by 0.01*.

Sample Size (Number of Pairs)

N (Sample Size)

Enter one or more values for the sample size. This is the number of pairs selected at random from the population to be in the study.

You can enter a single value or a range of values.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the interval to be used will be a one-sided or a two-sided confidence interval.

Precision

Distance from Mean Difference to Limit(s)

This is the distance from the confidence limit(s) to the mean paired difference. For two-sided intervals, it is also known as the precision, half-width, or margin of error.

You can enter a single value or a list of values. The value(s) must be greater than zero.

Standard Deviation of Paired Differences

S (Standard Deviation)

Enter a value (or range of values) for the standard deviation. You can use the results of a pilot study, a previous study, or a ball park estimate based on the range (e.g., Range/4) to estimate this parameter.

Know Standard Deviation

Check this box when you want to base your results on the normal distribution. When the box is not checked, calculations are based on the t-distribution. The difference between the two distributions is negligible when the sample sizes are large (>50).

Population

Population Size

This is the number of pairs in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made. This option sets the population size.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Sample Size

A researcher would like to estimate the mean difference in weight following a specific diet using a two-sided 95% confidence interval. The confidence level is set at 0.95, but 0.99 is included for comparative purposes. The standard deviation estimate, based on the range of paired differences, is 9.6 lbs. The researcher would like the interval to be no wider than 10 lbs. (half-width = 5 lbs.), but will examine half-widths of 3, 4, 5, 6, and 7 lbs.

The goal is to determine the necessary sample size.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for Paired Means** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Paired Means**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95 0.99**
N (Sample Size)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean to Limit(s)	**3 to 7 by 1**
S (Standard Deviation)	**9.6**
Population Size	**Infinite**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

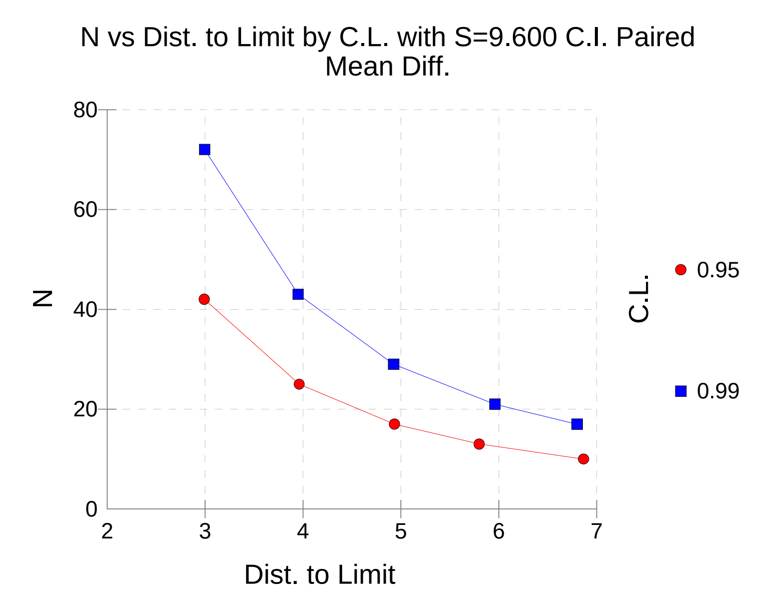
Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals with Unknown Standard Deviation

                                 Target       Actual
                   Sample     Dist from    Dist from     Standard
Confidence           Size     Mean Diff    Mean Diff    Deviation
Level                 (N)     to Limits    to Limits          (S)
0.95000                42         3.000        2.992        9.600
0.99000                72         3.000        2.995        9.600
0.95000                25         4.000        3.963        9.600
0.99000                43         4.000        3.950        9.600
0.95000                17         5.000        4.936        9.600
0.99000                29         5.000        4.926        9.600
0.95000                13         6.000        5.801        9.600
0.99000                21         6.000        5.961        9.600
0.95000                10         7.000        6.867        9.600
0.99000                17         7.000        6.801        9.600

References
Hahn, G. J. and Meeker, W.Q. 1991. Statistical Intervals. John Wiley & Sons. New York.
```

```
Report Definitions
Confidence level is the proportion of confidence intervals (constructed with this same confidence level,
    sample size, etc.) that would contain the population mean difference.
N is the size of the sample (or number of pairs) drawn from the population.
Dist from Mean Diff to Limit is the distance from the confidence limit(s) to the mean paired difference. For
    two-sided intervals, it is also know as the precision, half-width, or margin of error.
Target Dist from Mean Diff to Limit is the value of the distance that is entered into the procedure.
Actual Dist from Mean Diff to Limit is the value of the distance that is obtained from the procedure.
The standard deviation (S) is the standard deviation of the paired differences.

Summary Statements
A sample size of 42 produces a two-sided 95% confidence interval with a distance from the mean
paired difference to the limits that is equal to 2.992 when the estimated standard deviation of
the paired differences is 9.600.
```

This report shows the calculated sample size for each of the scenarios.

Plots Section



This plot shows the sample size versus the precision for the two confidence limits.

Example 2 – Validation

This procedure uses the same mechanics as the Confidence Intervals for One Mean procedure. The validation of this procedure is given in Examples 2 and 3 of the Confidence Intervals for One Mean procedure.

Chapter 497

Confidence Intervals for Paired Means with Tolerance Probability

Introduction

This routine calculates the sample size necessary to achieve a specified distance from the paired sample mean difference to the confidence limit(s) with a given tolerance probability at a stated confidence level for a confidence interval about a single mean difference when the underlying data distribution is normal.

Technical Details

For a paired sample mean difference from a normal distribution with unknown variance, a two-sided, 100(1 – α)% confidence interval is calculated by

$$\overline{X}_{Diff} \pm \frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{Diff}}{\sqrt{n}}$$

where $\overline{X}_{Diff}$ is the mean of the paired differences of the sample, and σ_{diff} is the known standard deviation of paired sample differences.

A one-sided 100(1 – α)% upper confidence limit is calculated by

$$\overline{X}_{Diff} + \frac{t_{1-\alpha,n-1}\hat{\sigma}_{Diff}}{\sqrt{n}}$$

Similarly, the one-sided 100(1 – α)% lower confidence limit is

$$\overline{X}_{Diff} - \frac{t_{1-\alpha,n-1}\hat{\sigma}_{Diff}}{\sqrt{n}}$$

Each confidence interval is calculated using an estimate of the mean difference plus and/or minus a quantity that represents the distance from the mean difference to the edge of the interval. For two-sided confidence intervals, this distance is sometimes called the precision, margin of error, or half-width. We will label this distance, *D*.

The basic equation for determining sample size when *D* has been specified is

$$D = \frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{Diff}}{\sqrt{n}}$$

Solving for *n*, we obtain

$$n = \left(\frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{Diff}}{D}\right)^2$$

This equation can be solved for any of the unknown quantities in terms of the others. The value $\alpha/2$ is replaced by α when a one-sided interval is used.

There is an additional subtlety that arises when the standard deviation is to be chosen for estimating sample size. The sample sizes determined from the formula above produce confidence intervals with the specified widths only when the future sample has a sample standard deviation of differences that is no greater than the value specified.

As an example, suppose that 15 pairs of individuals are sampled in a pilot study, and a standard deviation estimate of 3.5 is obtained from the sample. The purpose of a later study is to estimate the mean difference within 10 units. Suppose further that the sample size needed is calculated to be 57 pairs using the formula above with 3.5 as the estimate for the standard deviation. The sample of size 57 pairs is then obtained from the population, but the standard deviation of the 57 paired differences turns out to be 3.9 rather than 3.5. The confidence interval is computed and the distance from the mean difference to the confidence limits is greater than 10 units.

This example illustrates the need for an adjustment to adjust the sample size such that the distance from the mean difference to the confidence limits will be below the specified value with known probability.

Such an adjustment for situations where a previous sample is used to estimate the standard deviation is derived by Harris, Horvitz, and Mood (1948) and discussed in Zar (1984) and Hahn and Meeker (1991). The adjustment is

$$n = \left(\frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{Diff}}{D}\right)^2 F_{1-\gamma;n-1,m-1}$$

where $1 - \gamma$ is the probability that the distance from the mean difference to the confidence limit(s) will be below the specified value, and *m* is the sample size in the previous paired sample that was used to estimate the standard deviation.

The corresponding adjustment when no previous sample is available is discussed in Kupper and Hafner (1989) and Hahn and Meeker (1991). The adjustment in this case is

$$n = \left(\frac{t_{1-\alpha/2,n-1}\hat{\sigma}_{Diff}}{D}\right)^2 \left(\frac{\chi^2_{1-\gamma,n-1}}{n-1}\right)$$

where, again, $1 - \gamma$ is the probability that the distance from the mean difference to the confidence limit(s) will be below the specified value.

Each of these adjustments accounts for the variability in a future estimate of the standard deviation. In the first adjustment formula (Harris, Horvitz, and Mood, 1948), the distribution of the standard deviation is based on the estimate from a previous paired sample. In the second adjustment formula, the distribution of the standard deviation is based on a specified value that is assumed to be the population standard deviation of differences.

Finite Population Size

The above calculations assume that samples are being drawn from a large (infinite) population. When the population is of finite size (*N*), an adjustment must be made. The adjustment reduces the standard deviation as follows:

$$\sigma_{finite} = \sigma\sqrt{\left(1-\frac{n}{N}\right)}$$

This new standard deviation replaces the regular standard deviation in the above formulas.

Confidence Level

The confidence level, 1 – α, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is 1 – α.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters.

Confidence and Tolerance

Confidence Level (1 – Alpha)

The confidence level, 1 – α, has the following interpretation. If thousands of samples of *n* items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population mean difference is 1 – α.

Often, the values 0.95 or 0.99 are used. You can enter single values or a range of values such as *0.90 0.95 0.99* or *0.90 to 0.99 by 0.01*.

Tolerance Probability

This is the probability that a future interval with sample size *N* and the specified confidence level will have a distance from the mean paired difference to the limit(s) that is less than or equal to the distance specified.

If a tolerance probability is not used, as in the 'Confidence Intervals for Paired Means' procedure, the sample size is calculated for the expected distance from the mean paired difference to the limit(s), which assumes that the future standard deviation will also be the one specified.

Using a tolerance probability implies that the standard deviation of the future sample will not be known in advance, and therefore, an adjustment is made to the sample size formula to account for the variability in the standard deviation. Use of a tolerance probability is similar to using an upper bound for the standard deviation in the 'Confidence Intervals for Paired Means' procedure.

Values between 0 and 1 can be entered. The choice of the tolerance probability depends upon how important it is that the distance from the interval limit(s) to the mean difference is at most the value specified.

You can enter a range of values such as *0.70 0.80 0.90* or *0.70 to 0.95 by 0.05*.

Sample Size (Number of Pairs)

N (Sample Size or Number of Pairs)

Enter one or more values for the sample size. This is the number of pairs selected at random from the population to be in the study.

You can enter a single value or a range of values.

One-Sided or Two-Sided Interval

Interval Type

Specify whether the interval to be used will be a one-sided or a two-sided confidence interval.

Precision

Distance from Mean Difference to Limit(s)

This is the distance from the confidence limit(s) to the mean paired difference. For two-sided intervals, it is also known as the precision, half-width, or margin of error.

You can enter a single value or a list of values. The value(s) must be greater than zero.

Standard Deviation of Paired Differences

Standard Deviation Source

This procedure permits two sources for estimates of the standard deviation of paired differences:

- **S is a Population Standard Deviation**

  This option should be selected if there is no previous sample that can be used to obtain an estimate of the standard deviation of the paired differences. In this case, the algorithm assumes that future sample obtained will be from a population with standard deviation S.

- **S from a Previous Sample**

  This option should be selected if the estimate of the standard deviation of the paired differences is obtained from a previous random sample from the same distribution as the one to be sampled. The sample size of the previous sample must also be entered under 'Sample Size of Previous Sample'.

Standard Deviation of Paired Differences– S is a Population Standard Deviation

S (Standard Deviation)

Enter an estimate of the standard deviation of paired differences (must be positive). In this case, the algorithm assumes that future samples obtained will be from a population with standard deviation S.

One common method for estimating the standard deviation is the range divided by 4, 5, or 6.

You can enter a range of values such as *1 2 3* or *1 to 10 by 1*.

Press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Standard Deviation of Paired Differences – S from a Previous Sample

S (SD Estimated from a Previous Sample)

Enter an estimate of the standard deviation of paired differences from a previous (or pilot) study. This value must be positive.

A range of values may be entered.

Press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Sample Size (# of Pairs) of Previous Sample

Enter the sample size (number of pairs) that was used to estimate the standard deviation entered in S (SD Estimated from a Previous Sample).

This value is entered only when 'Standard Deviation Source:' is set to 'S from a Previous Sample'.

Population

Population Size

This is the number of pairs in the population. Usually, you assume that samples are drawn from a very large (infinite) population. Occasionally, however, situations arise in which the population of interest is of limited size. In these cases, appropriate adjustments must be made. This option sets the population size.

Iterations Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations allowed before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank. A value of 500 is recommended.

Example 1 – Calculating Sample Size

A researcher would like to estimate the mean difference in weight following a specific diet with 95% confidence. It is very important that the mean difference is estimated within 5 lbs. Data available from a previous study are used to provide an estimate of the standard deviation. The estimate of the standard deviation of before/after differences is 16.7 lbs, from a sample of size 17 individuals.

The goal is to determine the sample size necessary to obtain a two-sided confidence interval such that the mean weight is estimated within 5 lbs. Tolerance probabilities of 0.70 to 0.95 will be examined.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Confidence Intervals for Paired Means with Tolerance Probability** procedure window by clicking on **Confidence Intervals**, then **Means**, then **Paired Means with Tolerance Probability**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Confidence Level	**0.95**
Tolerance Probability	**0.70 to 0.95 by 0.05**
N (Sample Size or Number of Pairs)	*Ignored since this is the Find setting*
Interval Type	**Two-Sided**
Distance from Mean Diff to Limit(s)	**5**

Data Tab (continued)

Standard Deviation Source	**S from a Previous Sample**
S	**16.7**
Sample Size of Previous Sample	**17**
Population Size	**Infinite**

Annotated Output

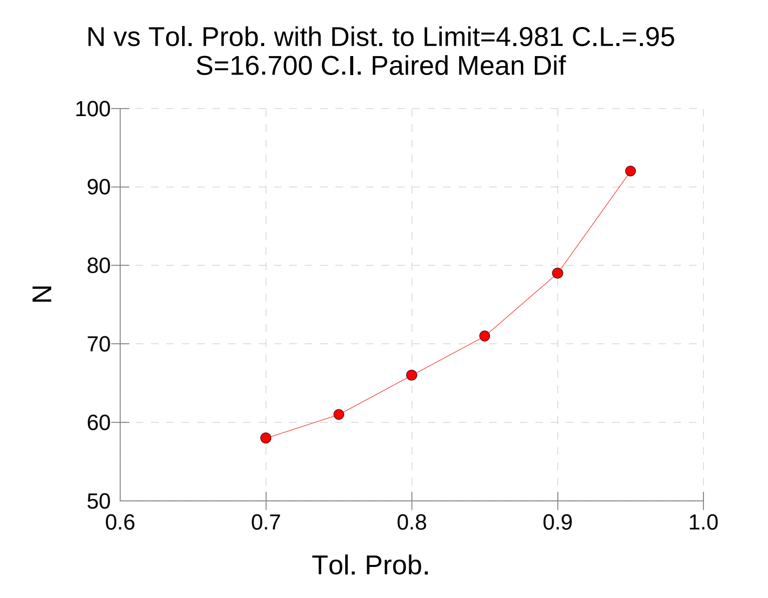
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Two-Sided Confidence Intervals

                                   Target        Actual
                     Sample     Dist from     Dist from      Standard
Confidence             Size     Mean Diff     Mean Diff     Deviation     Tolerance
Level                   (N)     to Limits     to Limits           (S)   Probability
0.95                     58         5.000         4.970        16.700          0.70
0.95                     61         5.000         4.996        16.700          0.75
0.95                     66         5.000         4.967        16.700          0.80
0.95                     71         5.000         4.985        16.700          0.85
0.95                     79         5.000         4.973        16.700          0.90
0.95                     92         5.000         4.981        16.700          0.95

Sample size for estimate of S from previous paired sample = 17.

References
Hahn, G. J. and Meeker, W.Q. 1991. Statistical Intervals. John Wiley & Sons. New York.
Zar, J. H. 1984. Biostatistical Analysis. Second Edition. Prentice-Hall. Englewood Cliffs, New Jersey.
Harris, M., Horvitz, D. J., and Mood, A. M. 1948. 'On the Determination of Sample Sizes in Designing
   Experiments', Journal of the American Statistical Association, Volume 43, No. 243, pp. 391-402.

Report Definitions
Confidence level is the proportion of confidence intervals (constructed with this same confidence level,
   sample size, etc.) that would contain the population mean difference.
N is the size of the sample (or number of pairs) drawn from the population.
Dist from Mean Diff to Limit is the distance from the confidence limit(s) to the mean paired difference. For
   two-sided intervals, it is also know as the precision, half-width, or margin of error.
Target Dist from Mean Diff to Limit is the value of the distance that is entered into the procedure.
Actual Dist from Mean Diff to Limit is the value of the distance that is obtained from the procedure.
The standard deviation (S) is the standard deviation of the paired differences.
Tolerance Probability is the probability that a future interval with sample size N and corresponding
   confidence level will have a distance from the mean difference to the limit(s) that is less than or equal to
   the specified distance.

Summary Statements
The probability is 0.70 that a sample size of 58 will produce a two-sided 95% confidence
interval with a distance from the mean paired difference to the limits that is less than or
equal to 4.970 if the population standard deviation is estimated to be 16.700 by a previous
paired sample of size 17.
```

This report shows the calculated sample size for each of the scenarios.

Plots Section



This plot shows the sample size versus the tolerance probability.

Example 2 – Validation

This procedure uses the same mechanics as the Confidence Intervals for One Mean with Tolerance Probability procedure. The validation of this procedure is given in Examples 2, 3, and 4 of the Confidence Intervals for One Mean with Tolerance Probability procedure.

Chapter 500

Inequality Tests for Two Means in a 2x2 Cross-Over Design using Differences

Introduction

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments and the objective is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to the other. It is assumed that there is a *washout* period between treatments during which the response returns back to its baseline value. If this does not occur, there is said to be a *carry-over* effect.

A 2x2 cross-over design contains to two *sequences* (treatment orderings) and two time periods (occasions). One sequence receives treatment A followed by treatment B. The other sequence receives B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carry over to the second. Thus, the groups in this design are defined by the sequence in which the drugs are administered, not by the treatments they receive. Indeed, higher-order cross-over designs have been used in which the same treatment is used at both occasions.

Cross-over designs are employed because, if the no-carryover assumption is met, treatment differences are measured within a subject rather than between subjects—making a more precise measurement. Examples of the situations that might use a cross-over design are the comparison of anti-inflammatory drugs in arthritis and the comparison of hypotensive agents in essential hypertension. In both of these cases, symptoms are expected to return to their usual baseline level shortly after the treatment is stopped.

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment into the study may be easier because each patient will receive both treatments. Finally, it is often more difficult to obtain a subject than to obtain a measurement.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. It may be difficult to separate the treatment effect from the period effect, the carry-over effect of the previous treatment, and the interaction between period and treatment.

The design cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it should not be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it is often more difficult to keep patients enrolled in the study. It is arguably simpler to measure a subject once than to obtain their measurement twice. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

Technical Details

The 2x2 crossover design may be described as follows. Randomly assign the subjects to one of two sequence groups so that there are N_1 subjects in sequence one and N_2 subjects in sequence two. In order to achieve design balance, the sample sizes N_1 and N_2 are assumed to be equal so that $N_1 = N_2 = N/2$.

Sequence one is given treatment A followed by treatment B. Sequence two is given treatment B followed by treatment A. The sequence is replicated *m* times. So, if *m* = 3, the sequences are ABABAB and BABABA.

The usual method of analysis is the analysis of variance. However, the power and sample size formulas that follow are based on the t-test, not the F-test. This is done because, in the balanced case, the t-test and the analysis of variance F-test are equivalent. Also, the F-test is limited to a two-sided hypothesis, while the t-test allows both one-sided and two-sided hypotheses. This is important because one-sided hypotheses are used for non-inferiority and equivalence testing.

Cross-Over Analysis

The following discussion summarizes the presentation of Chow and Liu (1999). The general linear model for the standard 2x2 cross-over design is

$$Y_{ijkl} = \mu + S_{ik} + P_j + \mu_{(j,k)} + C_{(j-1,k)} + e_{ijkl}$$

where *i* represents a subject (1 to N_k), *j* represents the period (1 or 2), *k* represents the sequence (1 or 2), and *l* represents the replicate. The S_{ik} represent the random effects of the subjects. The

P_j represent the effects of the two periods. The $\mu_{(j,k)}$ represent the means of the two treatments. In the case of the 2x2 cross-over design

$$\mu_{(j,k)} = \begin{cases} \mu_1 & if \;\; k = j \\ \mu_2 & if \;\; k \neq j \end{cases}$$

where the subscripts 1 and 2 represent treatments A and B, respectively.

The $C_{(j-1,k)}$ represent the carry-over effects. In the case of the 2x2 cross-over design

$$C_{(j-1,k)} = \begin{cases} C_1 & if \;\; j = 2, k = 1 \\ C_2 & if \;\; j = 2, k = 2 \\ 0 & otherwise \end{cases}$$

where the subscripts 1 and 2 represent treatments A and B, respectively.

Assuming that the average effect of the subjects is zero, the four means from the 2x2 cross-over design can be summarized using the following table.

$Sequence$	$Period\ 1$	$Period\ 2$
$1\,(AB)$	$\mu_{11} = \mu + P_1 + \mu_1$	$\mu_{21} = \mu + P_2 + \mu_2 + C_1$
$2\,(BA)$	$\mu_{12} = \mu + P_1 + \mu_2$	$\mu_{22} = \mu + P_2 + \mu_1 + C_2$

where $P_1 + P_2 = 0$ and $C_1 + C_2 = 0$.

Test Statistic

The presence of a treatment effect can be studied by testing whether $\mu_1 - \mu_2 = \delta$ using a *t*-test or an F-test. If the F-test is used, only a two-sided test is possible. The t statistic is calculated as follows

$$t_d = \frac{(\bar{x}_T - \bar{x}_R) - \delta}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

where $\hat{\sigma}_w^2$ is the within mean square error from the appropriate ANOVA table.

The two-sided null hypothesis is rejected at the α significance level if $|t_d| > t_{\alpha/2,N-2}$. Similar results are available for a one-sided hypothesis test.

The F-test is calculated using a standard repeated-measures analysis of variance table in which the between factor is the sequence and the within factor is the treatment. The within mean square error provides an estimate of the within-subject variance σ_w^2. If prior studies used a t-test rather than an ANOVA to analyze the data, you may not have a direct estimate of σ_w^2. Instead, you will have an estimate of the variance of the period differences from the t-test, $\hat{\sigma}_d^2$.The two variances, σ_d^2 and σ_w^2, are functionally related by $\sigma_w^2 = 2\sigma_d^2$. Either variance can be entered.

Computing the Power

The power is calculated as follows for a directional alternative (one-sided test).

1. Find t_α such that $1 - T_{df}\left(t_\alpha\right) = \alpha$, where $T_{df}\left(x\right)$ is the area left of *x* under a central-*t* curve and *df* = *N* - 2.

2. Calculate the noncentrality parameter: $\lambda = \dfrac{\delta\sqrt{N}}{\sigma_w\sqrt{2}}$.

3. Calculate: Power = $1 - T'_{df,\lambda}\left(t_\alpha\right)$, where $T'_{df,\lambda}\left(x\right)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power and Beta* or *N*.

Select *N* when you want to determine the sample size needed to achieve a given power and alpha level.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in both sequences). This value must be an integer greater than one.

Note that you may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Mean Differences

Diff0 (Mean Difference|H0)

Enter the difference between the treatment means under the null (H0) hypothesis. This is the value that is to be rejected when the t-test is significant. This value is commonly set to zero.

You may enter a range of values such as *10 20 30* or *0 to 100 by 25*.

Diff1 (Mean Difference|H1)

Enter the difference between the population means under the alternative (H1) hypothesis. This is the value of the difference at which the power is calculated.

You may enter a range of values such as *10 20 30* or *0 to 100 by 25*.

Effect Size – Standard Deviation

Specify S as Sw or Sd

Specify the form of the standard deviation that is entered in the box below.

- **Sw**

  Specify S as the square root of the within mean square error from a repeated measures ANOVA. This is the most common method since cross-over designs are usually analyzed using ANOVA.

- **Sd**

  Specify S as the standard deviation of the individual differences created for each subject. This option is used when you have previous studies that have produced this value.

S (Value of Sw or Sd)

Specify the value(s) of the standard deviation S. The interpretation of this value depends on the entry in *Specify S as Sw or Sd* above. If S=Sw is selected, this is the value of Sw which is

SQR(WMSE) where WMSE is the within mean square error from the ANOVA table used to analyze the Cross-Over design. If S = Sd is selected, this is the value of Sd which is the standard deviation of the period differences—pooled from both sequences.

These values must be positive. A list of values may be entered.

You can press the SD button to load the Standard Deviation Estimator window.

Test

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. The null hypothesis is always H_0: Diff0 = Diff1.

Note that the alternative hypothesis enters into power calculations by specifying the rejection region of the hypothesis test. Its accuracy is critical.

Possible selections are:

- **H1: DIFF0 <> DIFF1**

  This is the most common selection. It yields the *two-sided* t-test. Use this option when you are testing whether the means are different but you do not want to specify beforehand which mean is larger. Many scientific journals require two-sided tests.

- **H1: DIFF0 > DIFF1**

  This option yields a *one-sided* t-test. Use it when you are only interested in the case in which the actual difference is less than Diff0.

- **H1: DIFF0 < DIFF1**

  This option yields a *one-sided* t-test. Use it when you are only interested in the case in which actual difference is greater than Diff0.

Example 1 – Power Analysis

Suppose you want to consider the power of a balanced cross-over design that will be analyzed using the two-sided t-test approach. The difference between the treatment means under H0 is 0. Similar experiments have had a standard deviation of the differences (Sd) of 10. Compute the power when the true differences are 5 and 10 at sample sizes between 5 and 50. The significance level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Inequality Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**5 10 15 20 30 40 50**
Diff0 (Mean Difference|H0)	**0**
Diff1 (Mean Difference|H1)	**5 10**
Specify S as Sw or Sd	**Sd**
S (Value of Sw or Sd)	**10**
Alternative Hypothesis	**H1: Diff0 <> Diff1**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

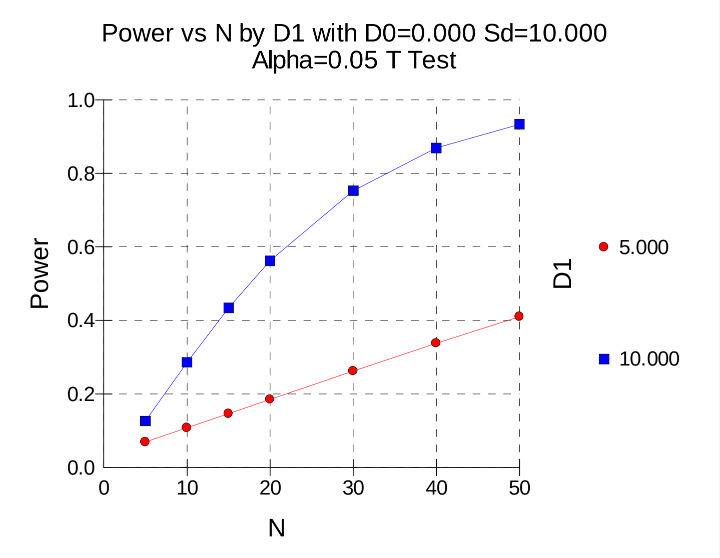
Numeric Results

```
Numeric Results for Cross-Over Design
Null Hypothesis: Diff0=Diff1    Alternative Hypothesis: Diff0<>Diff1

                                                                   Effect
Power        N      Diff0      Diff1    Alpha     Beta       Sd      Size
0.0691       5      0.000      5.000   0.0500   0.9309   10.000     0.500
0.1077      10      0.000      5.000   0.0500   0.8923   10.000     0.500
0.1463      15      0.000      5.000   0.0500   0.8537   10.000     0.500
0.1851      20      0.000      5.000   0.0500   0.8149   10.000     0.500
0.2624      30      0.000      5.000   0.0500   0.7376   10.000     0.500
0.3379      40      0.000      5.000   0.0500   0.6621   10.000     0.500
0.4101      50      0.000      5.000   0.0500   0.5899   10.000     0.500
0.1266       5      0.000     10.000   0.0500   0.8734   10.000     1.000
0.2863      10      0.000     10.000   0.0500   0.7137   10.000     1.000
0.4339      15      0.000     10.000   0.0500   0.5661   10.000     1.000
0.5620      20      0.000     10.000   0.0500   0.4380   10.000     1.000
0.7529      30      0.000     10.000   0.0500   0.2471   10.000     1.000
0.8690      40      0.000     10.000   0.0500   0.1310   10.000     1.000
0.9337      50      0.000     10.000   0.0500   0.0663   10.000     1.000
```

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the total sample size drawn from all sequences. The sample is divided equally among sequences.
Alpha is the probability of a false positive.
Beta is the probability of a false negative.
Diff0 is the mean difference under the null hypothesis, H0.
Diff1 is the mean difference under the alternative hypothesis, H1.
Sd is the standard deviation of the difference.
Effect Size, |Diff0-Diff1|/Sd, is the relative magnitude of the effect under the alternative.

Summary Statements
A two-sided t-test achieves 7% power to infer that the mean difference is not 0.000 when the total sample size of a 2x2 cross-over design is 5, the actual mean difference is 5.000, the standard deviation of the differences is 10.000, and the significance level is 0.0500.

Plots Sections



This report shows the values of each of the parameters, one scenario per row. This plot shows the relationship between sample size and power. We see that a sample size of about 46 is needed when Diff1 = 10 for 90% power, while Diff1 = 5 never reaches 90% power in this range of sample sizes.

Example 2 – Finding the Sample Size

Continuing with Example 1, suppose the researchers want to find the exact sample size necessary to achieve 90% power for both values of Diff1.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Inequality Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power	**0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
Diff0 (Mean Difference|H0)	**0**
Diff1 (Mean Difference|H1)	**5 10**
Specify S as Sw or Sd	**Sd**
S (Value of Sw or Sd)	**10**
Alternative Hypothesis	**H1: Diff0 <> Diff1**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Cross-Over Design
Null Hypothesis: Diff0=Diff1     Alternative Hypothesis: Diff0<>Diff1

                                                                       Effect
Power          N        Diff0        Diff1     Alpha      Beta        Sd     Size
0.9032       172        0.000        5.000    0.0500    0.0968    10.000    0.500
0.9125        46        0.000       10.000    0.0500    0.0875    10.000    1.000
```

This report shows the exact sample size necessary for each scenario.

Note that the search for N is conducted across only even values of N since the design is assumed to be balanced.

Example 3 – Validation using Julious

Julious (2004) page 1933 presents an example in which Diff0 = 0.0, Diff1 = 10, Sw = 20, alpha = 0.05, and beta = 0.10. Julious obtains a sample size of 86.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Inequality Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power	**0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
Diff0 (Mean Difference|H0)	**0**
Diff1 (Mean Difference|H1)	**10**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**20**
Alternative Hypothesis	**H1: Diff0 <> Diff1**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                                                                          Effect
Power        N     Diff0     Diff1      Alpha        Beta         Sw      Size
0.906483     88    0.000     10.000     0.050000     0.093435     20.000  0.500
```

PASS obtained a sample size of 88, two higher than that obtained by Julious (2004). However, if you look at the power achieved by an N of 86, you will find that it is 0.899997—slightly less than the goal of 0.90.

Chapter 505

Inequality Tests for Two Means in a 2x2 Cross-Over Design using Ratios

Introduction

This procedure calculates power and sample size for a 2x2 cross-over design in which the logarithm of the outcome is a continuous normal random variable. This routine deals with the case in which the statistical hypotheses are expressed in terms of ratios of means instead of differences of means.

The details of testing two treatments using data from a 2x2 cross-over design are given in another chapter and they will not be repeated here. If the logarithms of the responses can be assumed to follow a normal distribution, hypotheses stated in terms of the ratio can be transformed into hypotheses about the difference. The details of this analysis are given in Julious (2004). They will only be summarized here.

Testing Using Ratios

It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment (group 2) mean.
μ_R	Not used	*Reference mean*. This is the reference (group 1) mean.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only the ratio of these values is needed for power and sample size calculations.

The null hypothesis is

$$H_0: \phi = \phi_0$$

and the alternative hypothesis is

$$H_1: \phi \neq \phi_0$$

Log Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of ratios.
2. Transform these into hypotheses about differences by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the null hypothesis are as follows.

$$\begin{aligned}
&\phi = \phi_0 \\
&\Rightarrow \phi = \left\{\frac{\mu_T}{\mu_R}\right\} \\
&\Rightarrow \ln(\phi) \neq \left\{\ln(\mu_T) - \ln(\mu_R)\right\}
\end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter can be used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable X is the logarithm of the original variable Y. That is, $X = \ln(Y)$ and $Y = \exp(X)$. Label the mean and variance of X as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of Y as μ_Y and σ_Y^2, respectively. If X is normally distributed, then Y is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be found to be

$$\begin{aligned} COV_Y &= \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y} \\ &= \sqrt{e^{\sigma_X^2} - 1} \\ &= \sqrt{e^{\sigma_w^2} - 1} \end{aligned}$$

where σ_w^2 is the within mean square error from the analysis of variance of the logged data. Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2 + 1 \right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

Thus, the hypotheses can be stated in the original (Y) scale and then the power can be analyzed in the transformed (X) scale.

Power Calculation

As is shown above, the hypotheses can be stated in the original (Y) scale using ratios or the logged (X) scale using differences. In either case, the power and sample size calculations are made using the formulas for testing the difference in two means. These formulas are presented in another chapter and are not duplicated here.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. In most situations, you will select either *Power and Beta* for a power analysis or *N1* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the total sample size, the number of individuals in the study (total subjects in both sequences). This value must be an integer greater than one.

When N is even, it is split evenly between the two sequences. When N is odd, the first sequence has one more subject than the second sequence.

Note that you may enter a list of values using the syntax *50 100 150 200 250* or *50 to 250 by 50*.

Effect Size – Ratios

R0 (Ratio Under H0)

This is the value of the ratio of the two means assumed by the null hypothesis, H0. Usually, R0 = 1.0 which implies that the two means are equal. However, you may test other values of R0 as well. Strictly speaking, any positive number is valid, but, usually, 1.0 is used.

Warning: you cannot use the same value for both R0 and R1.

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Often, a range of values will be tried. For example, you might try the four values:

1.05 1.10 1.15 1.20

Strictly speaking, any positive number is valid. However, numbers between 0.50 and 2.00 are usually used.

Warning: you cannot use the same value for both R0 and R1.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not log) scale. This value must be determined from past experience or from a pilot study. It is most easily calculated from the within mean-square error of the analysis of variance of the logged data using the relationship

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\ .$$

If prior studies used a t-test to analyze the logged data, you will not have a direct estimate of $\hat{\sigma}_w^2$. However, the two variances, σ_d^2 and σ_w^2, are functionally related by $\sigma_d^2 = 2\sigma_w^2$.

Test

Alternative Hypothesis

This option specifies the alternative hypothesis. This implicitly specifies the direction of the hypothesis test. Possible selections are:

- **H1: R1 <> R0**

  This is the most common selection. It yields the *two-tailed t-test*. Use this option when you are testing whether the means are different, but you do not want to specify beforehand which mean is larger.

- **H1: R1 < R0**

  This option yields a *one-tailed t-test*. Use it when you are only interested in the case in which *Mean1* is greater than *Mean2*.

- **H1: R1 > R0**

  This option yields a *one-tailed t-test*. Use it when you are only interested in the case in which *Mean1* is less than *Mean2*.

Example 1 – Finding Power

A company has developed a generic drug for treating rheumatism and wants to show that it is better than the standard drug. Responses for either treatment are assumed to follow a lognormal distribution. A 2x2 cross-over design will be used and the logged data will be analyzed using an appropriate analysis of variance. Note that using an analysis of variance instead of a t-test to analyze the data forces the researchers to use two-sided tests.

Past experience leads the researchers to set the COV to 0.50. The significance level is 0.05. The power will be computed for R1 equal to 1.10 and 1.20. Sample sizes between 20 and 220 will be included in the initial analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a 2x2 Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Inequality Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**20 to 220 by 40**
R0 (Ratio Under H0)	**1.0**
R1 (True Ratio)	**1.1 1.2**
COV (Coefficient of Variation)	**0.50**
Alternative Hypothesis	**R1<>R0 (Two-Sided)**

Annotated Output

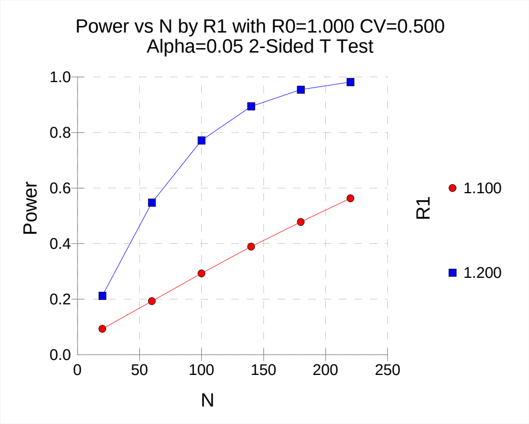
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for 2x2 Cross-Over Design Using Ratios
H0: R1=R0.  H1: R1<>R0.

                 Total        Mean         Mean                 Coefficient
                Sample       Ratio        Ratio      Effect            of    Significance
                  Size    Under H0     Under H1        Size     Variation           Level
Power              (N)        (R0)         (R1)        (ES)         (COV)         (Alpha)        Beta
0.0928              20       1.000        1.100       0.143         0.500          0.0500      0.9072
0.1925              60       1.000        1.100       0.143         0.500          0.0500      0.8075
0.2925             100       1.000        1.100       0.143         0.500          0.0500      0.7075
0.3885             140       1.000        1.100       0.143         0.500          0.0500      0.6115
0.4777             180       1.000        1.100       0.143         0.500          0.0500      0.5223
0.5627             220       1.000        1.100       0.143         0.500          0.0500      0.4373
0.2116              20       1.000        1.200       0.273         0.500          0.0500      0.7884
0.5474              60       1.000        1.200       0.273         0.500          0.0500      0.4526
0.7711             100       1.000        1.200       0.273         0.500          0.0500      0.2289
0.8937             140       1.000        1.200       0.273         0.500          0.0500      0.1063
0.9537             180       1.000        1.200       0.273         0.500          0.0500      0.0463
0.9808             220       1.000        1.200       0.273         0.500          0.0500      0.0192

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the total sample size drawn from all sequences. The sample is divided equally among sequences.
R0 is the ratio of the means (Mean2/Mean1) under the null hypothesis, H0.
R1 is the ratio of the means (Mean2/Mean1) at which the power is calculated.
ES is the effect size which is |Ln(R0)-Ln(R1)| / (sigma).
COV is the coefficient of variation on the original scale. The value of sigma is calculated from this.
Alpha is the probability of a false positive H0.
Beta is the probability of a false negative H0.

Summary Statements
A two-sided t-test achieves 9% power to infer that the mean ratio is not 1.000 when the total
sample size of a 2x2 cross-over design is 20, the actual mean ratio is 1.100, the coefficient
of variation is 0.500, and the significance level is 0.0500.
```

This report shows the power for the indicated scenarios.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Validation

We will validate this procedure by showing that it gives the identical results to the regular test on differences—a procedure that has been validated. We will use the same settings as those given in Example 1. Since the output for this example is shown above, all that we need is the output from the procedure that uses differences.

To run the power analysis on differences, we need the values of Diff1 (which correspond to R1) and Sw. The value of Diff0 will be zero.

$$\begin{aligned} Sw &= \sqrt{\ln\left(COV^2+1\right)} \\ &= \sqrt{\ln\left(0.5^2+1\right)} \\ &= 0.472381 \end{aligned}$$

$$\begin{aligned} Diff1 &= \ln\left(R1\right) & \qquad Diff1 &= \ln\left(R1\right) \\ &= \ln\left(1.10\right) & &= \ln\left(1.20\right) \\ &= 0.095310 & &= 0.182322 \end{aligned}$$

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Inequality Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Inequality Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1d** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**20 to 220 by 40**
Diff0 (Mean Difference|H0)	**0**
Diff1 (Mean Difference|H1)	**0.095310 0.182322**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**0.472381**
Alternative Hypothesis	**H1: DIFF0<>Diff1**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for 2x2 Cross-Over Design
Null Hypothesis: Diff0=Diff1     Alternative Hypothesis: Diff0<>Diff1
                                                                     Effect
Power          N     Diff0      Diff1     Alpha      Beta       Sw     Size
0.0928        20     0.000      0.095    0.0500    0.9072    0.472    0.202
0.1925        60     0.000      0.095    0.0500    0.8075    0.472    0.202
0.2925       100     0.000      0.095    0.0500    0.7075    0.472    0.202
0.3885       140     0.000      0.095    0.0500    0.6115    0.472    0.202
0.4777       180     0.000      0.095    0.0500    0.5223    0.472    0.202
0.5627       220     0.000      0.095    0.0500    0.4373    0.472    0.202
0.2116        20     0.000      0.182    0.0500    0.7884    0.472    0.386
0.5474        60     0.000      0.182    0.0500    0.4526    0.472    0.386
0.7711       100     0.000      0.182    0.0500    0.2289    0.472    0.386
0.8937       140     0.000      0.182    0.0500    0.1063    0.472    0.386
0.9537       180     0.000      0.182    0.0500    0.0463    0.472    0.386
0.9808       220     0.000      0.182    0.0500    0.0192    0.472    0.386
```

You can compare these power values with those shown above in Example 1 to validate the procedure. You will find that the power values are identical.

Chapter 510

Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design using Differences

Introduction

This procedure computes power and sample size for non-inferiority and superiority tests in 2x2 cross-over designs in which the outcome is a continuous normal random variable. The details of sample size calculation for the 2x2 cross-over design are presented in the 2x2 Cross-Over Designs chapter and they will not be duplicated here. This chapter only discusses those changes necessary for non-inferiority and superiority tests. Sample size formulas for non-inferiority and superiority tests of cross-over designs are presented in Chow et al. (2003) pages 63-68.

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments and the objective is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to the other. It is assumed that there is a *washout* period between treatments during which the response returns back to its baseline value. If this does not occur, there is said to be a *carry-over* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence receives treatment A followed by treatment B. The other sequence receives B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carry-over to the second. Thus, the groups in this design are defined by the sequence in which the two drugs are administered, not by the treatments they receive.

Cross-over designs are employed because, if the no-carryover assumption is met, treatment differences are measured within a subject rather than between subjects—making a more precise measurement. Examples of the situations that might use a cross-over design are the comparison of anti-inflammatory drugs in arthritis and the comparison of hypotensive agents in essential hypertension. In both of these cases, symptoms are expected to return to their usual baseline level shortly after the treatment is stopped.

The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests and their power and sample size can be calculated using the 2x2 Cross-Over Design procedure. However, at the urging of our users, we have developed this module which provides the input and output in formats that are convenient for these types of tests. This section reviews the specifics of non-inferiority and superiority testing.

Remember that in the usual t-test setting, the null (H0) and alternative (H1) hypotheses for one-sided tests are defined as

$$\mathrm{H}_0\!: \mu_X \le A \ \text{ versus } \ \mathrm{H}_1\!: \mu_X > A$$

Rejecting H0 implies that the mean is larger than the value *A*. This test is called an *upper-tailed test* because it is rejected in samples in which the difference in sample means is larger than *A.*

Following is an example of a *lower-tailed test*.

$$\mathrm{H}_0\!: \mu_X \ge A \ \text{ versus } \ \mathrm{H}_1\!: \mu_X < A$$

Non-inferiority and *superiority* tests are special cases of the above directional tests. It will be convenient to adopt the following specialize notation for the discussion of these tests.

Parameter	**PASS Input/Output**	**Interpretation**
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	|E|	*Margin of equivalence*. This is a tolerance value that defines the maximum difference that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value is shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	D	*True difference*. This is the value of $\mu_T - \mu_R$, the difference between the treatment and reference means. This is the value at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their difference is needed for power and sample size calculations.

Non-Inferiority and Superiority Tests

A *non-inferiority test* tests that the treatment mean is not worse than the reference mean by more than a small equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

A *superiority test* tests that the treatment mean is better than reference mean by more than a small equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

Case 1: High Values Good, Non-Inferiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no less than a small amount below the reference mean. The value of δ is often set to zero.

$$\mathrm{H_0}: \mu_T \le \mu_R - |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \mu_T > \mu_R - |\varepsilon|$$

$$\mathrm{H_0}: \mu_T - \mu_R \le -|\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \mu_T - \mu_R > -|\varepsilon|$$

$$\mathrm{H_0}: \delta \le -|\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \delta > -|\varepsilon|$$

Case 2: High Values Bad, Non-Inferiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no more than a small amount above the reference mean. The value of δ is often set to zero.

$$\mathrm{H_0}: \mu_T \ge \mu_R + |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \mu_T < \mu_R + |\varepsilon|$$

$$\mathrm{H_0}: \mu_T - \mu_R \ge |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \mu_T - \mu_R < |\varepsilon|$$

$$\mathrm{H_0}: \delta \ge |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \delta < |\varepsilon|$$

Case 3: High Values Good, Superiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is greater than the reference mean by at least the margin of equivalence. The value of δ must be greater than $|\varepsilon|$.

$$\mathrm{H_0}: \mu_T \le \mu_R + |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \mu_T > \mu_R + |\varepsilon|$$

$$\mathrm{H_0}: \mu_T - \mu_R \le |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \mu_T - \mu_R > |\varepsilon|$$

$$\mathrm{H_0}: \delta \le |\varepsilon| \quad \text{versus} \quad \mathrm{H_1}: \delta > |\varepsilon|$$

Case 4: High Values Bad, Superiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is less than the reference mean by at least the margin of equivalence. The value of δ must be less than $-|\varepsilon|$.

$\mathrm{H}_0{:}\,\mu_T \geq \mu_R - |\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T < \mu_R - |\varepsilon|$

$\mathrm{H}_0{:}\,\mu_T - \mu_R \geq -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\mu_T - \mu_R < -|\varepsilon|$

$\mathrm{H}_0{:}\,\delta \geq -|\varepsilon|$ versus $\mathrm{H}_1{:}\,\delta < -|\varepsilon|$

Test Statistics

This section describes the test statistic that is used to perform the hypothesis test.

T-Test

A t-test is used to analyze the data. When the data are balanced between sequences, the two-sided t-test is equivalent to an analysis of variance F-test. The test assumes that the data are a simple random sample from a population of normally-distributed values that have the same variance. This assumption implies that the differences are continuous and normal. The calculation of the t-statistic proceeds as follow

$$t_d = \frac{(\bar{x}_T - \bar{x}_R) - \varepsilon}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

where $\hat{\sigma}_w^2$ is the within mean square error from the appropriate ANOVA table.

The significance of the test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the hypothesis is rejected. That is, the one-sided null hypothesis is rejected at the α significance level if $t_d > t_{\alpha,N-2}$. Otherwise, no conclusion can be reached.

If prior studies used a t-test rather than an ANOVA to analyze the data, you may not have a direct estimate of σ_w^2. Instead, you will have an estimate of the variance of the period differences from the t-test, $\hat{\sigma}_d^2$. These variances are functionally related by $\sigma_w^2 = 2\sigma_d^2$. Either variance can be entered.

Computing the Power

The power is calculated as follows.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(x)$ is the area under a central-*t* curve to the left of *x* and *df* = *N* - 2.

2. Calculate the noncentrality parameter: $\lambda = \dfrac{(\delta - \varepsilon)\sqrt{N}}{\sigma_w \sqrt{2}}$.

3. Calculate: Power = $1 - T'_{df,\lambda}\left(t_{\alpha}\right)$, where $T'_{df,\lambda}\left(x\right)$ is the area under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ to the left of *x*.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power and Beta* or *N*.

Select *N* when you want to determine the sample size needed to achieve a given power and alpha level.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of inferiority when in fact the treatment mean is non-inferior.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of different means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study. This value must be an integer greater than one. Note that you may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Mean Difference

|E| (Equivalence Margin)

This is the magnitude of the *margin of equivalence*. It is the smallest difference between the mean and the reference value that still results in the conclusion of non-inferiority (or superiority). Note that the sign of this value is assigned depending on the selections for Higher Is and Test Type.

D (True Value)

This is the actual difference between the mean and the reference value. For non-inferiority tests, this value is often set to zero, but it can be non-zero as long as the values are consistent with the alternative hypothesis, H1. For superiority tests, this value is usually non-zero. Again, it must be consistent with the alternative hypothesis, H1.

Effect Size – Standard Deviation

Specify S as Sw or Sd

Specify the form of the standard deviation that is entered in the box below.

- **Sw**

  Specify the standard deviation S as the square root of the within mean square error from a repeated measures ANOVA. This is the most common method since cross-over designs are usually analyzed using ANOVA.

- **Sd**

  Specify the standard deviation S as the standard deviation of the individual treatment differences. This option is used when you have previous studies that produced this value.

S (Value of Sw or Sd)

Specify the value(s) of the standard deviation S. The interpretation of this value depends on the entry in *Specify S as Sw or Sd* above. If S=Sw is selected, this is the value of Sw which is SQR(WMSE) where WMSE is the within mean square error from the ANOVA table used to analyze the Cross-Over design. If S = Sd is selected, this is the value of Sd which is the standard deviation of the period differences—pooled from both sequences.

These values must be positive. A list of values may be entered.

You can press the SD button to load the Standard Deviation Estimator window.

Test

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the treatment mean is within the margin of equivalence of being no worse than the reference mean. Select *Superiority* when you want to test whether the treatment mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are usually considered bad. However, if the response variable is income, higher values are considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Example 1 – Power Analysis

Suppose you want to consider the power of a balanced, cross-over design that will be analyzed using the t-test approach. You want to compute the power when the margin of equivalence is either 5 or 10 at several sample sizes between 5 and 50. The true difference between the means under H0 is assumed to be 0. Similar experiments have had an *Sw* of 10. The significance level is 0.025.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

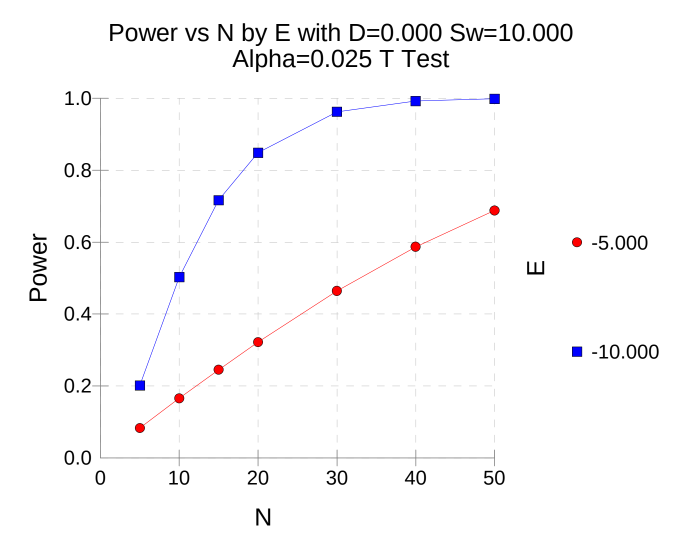
Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N (Total Sample Size)	**5 10 15 20 30 40 50**
|E| (Equivalence Margin)	**5 10**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**10**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Numeric Results for Non-Inferiority T-Test (H0: D <= -|E|; H1: D > -|E|)

                   Equivalence      Actual  Significance                       Standard
                        Margin  Difference         Level                      Deviation
Power          N           (E)         (D)       (Alpha)          Beta             (Sw)
0.08310        5        -5.000       0.000       0.02500       0.91690           10.000
0.16563       10        -5.000       0.000       0.02500       0.83437           10.000
0.24493       15        -5.000       0.000       0.02500       0.75507           10.000
0.32175       20        -5.000       0.000       0.02500       0.67825           10.000
0.46414       30        -5.000       0.000       0.02500       0.53586           10.000
0.58682       40        -5.000       0.000       0.02500       0.41318           10.000
0.68785       50        -5.000       0.000       0.02500       0.31215           10.000
0.20131        5       -10.000       0.000       0.02500       0.79869           10.000
0.50245       10       -10.000       0.000       0.02500       0.49755           10.000
0.71650       15       -10.000       0.000       0.02500       0.28350           10.000
0.84845       20       -10.000       0.000       0.02500       0.15155           10.000
0.96222       30       -10.000       0.000       0.02500       0.03778           10.000
0.99173       40       -10.000       0.000       0.02500       0.00827           10.000
0.99835       50       -10.000       0.000       0.02500       0.00165           10.000

Report Definitions
H0 (null hypothesis) is that D <= -|E|, where D = Treatment Mean - Reference Mean.
H1 (alternative hypothesis) is that D > -|E|.
Power is the probability of rejecting H0 when it is false. It should be close to one.
N is the total sample size drawn from all sequences. The sample is divided equally among sequences.
Alpha is the probability of a false positive H0.
Beta is the probability of a false negative H0.
|E| is the magnitude of the margin of equivalence. It is the largest difference that is not of practical significance.
D is actual difference between the treatment and reference means.
Sw is the square root of the within mean square error from the ANOVA table.

Summary Statements
A total sample size of 5 achieves 8% power to detect non-inferiority using a one-sided t-test
when the margin of equivalence is -5.000, the true mean difference is 0.000, the significance
level is 0.02500, and the square root of the within mean square error is 10.000. A 2x2
cross-over design with an equal number in each sequence is used.
```



This report shows the values of each of the parameters, one scenario per row. The plot shows the relationship between sample size and power. We see that a sample size of about 20 is needed to achieve 80% power when E = -10.

Example 2 – Finding the Sample Size

Continuing with Example 1, suppose the researchers want to find the exact sample size necessary to achieve 90% power for both values of D.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power	**0.90**
Alpha	**0.025**
N (Total Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**5 10**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**10**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority T-Test (H0: D <= -|E|; H1: D > -|E|)

                Equivalence     Actual  Significance                Standard
                     Margin Difference         Level                Deviation
Power        N          (E)        (D)       (Alpha)        Beta        (Sw)
0.90648     88       -5.000      0.000       0.02500     0.09352      10.000
0.91139     24      -10.000      0.000       0.02500     0.08861      10.000
```

This report shows the exact sample size necessary for each scenario.

Note that the search for N is conducted across only even values of N since the design is assumed to be balanced.

Example 3 – Validation using Julious

Julious (2004) page 1953 presents an example in which D = 0.0, E = 10, Sw = 20.00, alpha = 0.025, and beta = 0.10. Julious obtains a sample size of 86.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power	**0.10**
Alpha	**0.025**
N (Total Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**10**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**20**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority T-Test (H0: D <= -|E|; H1: D > -|E|)

                    Equivalence     Actual   Significance                 Standard
                         Margin Difference          Level                Deviation
Power           N           (E)        (D)        (Alpha)        Beta         (Sw)
0.90648        88       -10.000      0.000        0.02500     0.09352       20.000
```

PASS obtained a sample size of 88, two higher than that obtained by Julious (2004). However, if you look at the power achieved by an N of 86, you will find that it is 0.899997—slightly less than the goal of 0.90.

Chapter 515

Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design using Ratios

Introduction

This procedure calculates power and sample size of statistical tests for non-inferiority and superiority tests from a 2x2 cross-over design. This routine deals with the case in which the statistical hypotheses are expressed in terms mean ratios rather than mean differences.

The details of testing the non-inferiority of two treatments using data from a 2x2 cross-over design are given in another chapter and they will not be repeated here. If the logarithms of the responses can be assumed to follow the normal distribution, hypotheses about non-inferiority and superiority stated in terms of the ratio can be transformed into hypotheses about the difference. The details of this analysis are given in Julious (2004). They will only be summarized here.

Non-Inferiority Testing Using Ratios

It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	E	*Margin of equivalence.* This is a tolerance value that defines the maximum amount that is not of practical importance. This is the largest change in the mean ratio from the baseline value (usually one) that is still considered to be trivial.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their ratio is needed for power and sample size calculations.

The null hypothesis of inferiority is

$$\mathrm{H}_0{:}\phi \le \phi_L \quad where\ \phi_L < 1.$$

and the alternative hypothesis of non-inferiority is

$$\mathrm{H}_1{:}\phi > \phi_L$$

Log Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as scale-less percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of ratios.
2. Transform these into hypotheses about differences by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the null hypothesis are as follows.

$$\begin{aligned}
&\phi_L \le \phi \\
&\Rightarrow \phi_L \le \left\{\frac{\mu_T}{\mu_R}\right\} \\
&\Rightarrow \ln(\phi_L) \le \{\ln(\mu_T) - \ln(\mu_R)\}
\end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter is used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable *X* is the logarithm of the original variable *Y*. That is, *X* = ln(*Y*) and *Y* = exp(*X*). Label the mean and variance of *X* as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of *Y* as μ_Y and σ_Y^2, respectively. If *X* is normally distributed, then *Y* is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be expressed as

$$\begin{aligned} COV_Y &= \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y} \\ &= \sqrt{e^{\sigma_X^2} - 1} \\ &= \sqrt{e^{\sigma_w^2} - 1} \end{aligned}$$

where σ_w^2 is the within mean square error from the analysis of variance of the logged data. Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2 + 1 \right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

Thus, the hypotheses can be stated in the original (Y) scale and then power can be analyzed in the transformed (X) scale.

Power Calculation

As is shown above, the hypotheses can be stated in the original (Y) scale using ratios or the logged (X) scale using differences. Either way, the power and sample size calculations are made using the formulas for testing the equivalence of the difference in two means. These formulas are presented in another chapter and are not duplicated here.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for the probability of a type-II error (beta). A type-II error occurs when you fail to reject the null hypothesis of inferiority when in fact the treatment mean is non-inferior.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Recently, the standard has shifted to 0.10.

Power is defined as one minus beta. Power is equal to the probability of rejecting a false null hypothesis. Hence, specifying the beta error level also specifies the power level. For example, if you specify beta values of 0.05, 0.10, and 0.20, you are specifying the corresponding power values of 0.95, 0.90, and 0.80, respectively.

Alpha (Significance Level)

Specify one or more values of alpha. Alpha is the probably of a type-I error. A type-I error occurs when you reject the null hypothesis of inferiority when in fact the treatment group is not inferior to the reference group.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in both sequences). This value must be an integer greater than one.

When N is even, it is split evenly between the two sequences. When N is odd, the first sequence has one more subject than the second sequence.

Note that you may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Ratios

E (Equivalence Margin)

This is the magnitude of the relative *margin of equivalence*. It is the smallest change in the ratio of the two means that still results in the conclusion of non-inferiority (or superiority).

For example, suppose the non-inferiority boundary for the mean ratio is to be 0.80. This value is interpreted as follows: if the mean ratio (Treatment Mean / Reference Mean) is greater than 0.80, the treatment group is non-inferior to the reference group. In this example, the margin of equivalence would be 1.00 - 0.80 = 0.20.

This example assumed that higher values are better. If higher values are worse, an equivalence margin of 0.20 would be translated into a non-inferiority bound of 1.20. In this case, if the mean ratio is less than 1.20, the treatment group is non-inferior to the reference group.

Note that the sign of this value is ignored. Only the magnitude is used.

Recommended values:

0.20 is a common value for the parameter.

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Usually, the ratio will be assumed to be one. However, some authors recommend calculating the power using a ratio of 0.95 since this will require a larger sample size.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not logged) scale. This value must be determined from past experience or from a pilot study. It is most easily calculated from the within mean-square error of the analysis of variance of the logged data using the relationship

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\,.$$

If prior studies used a t-test to analyze the logged data, you will not have a direct estimate of $\hat{\sigma}_w^2$. However, the two variances, σ_d^2 and σ_w^2, are functionally related. The relationship between these quantities is $\sigma_d^2 = 2\sigma_w^2$.

Test

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the treatment mean is within the margin of equivalence of being no worse than the reference mean.

Select *Superiority* when you want to test whether the treatment mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are usually considered bad. However, if the response variable is income, higher values are probably considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Example 1 – Finding Power

A company has developed a generic drug for treating rheumatism and wants to show that it is not inferior to standard drug. A 2x2 cross-over design will be used to test the non-inferiority of the treatment drug to the reference drug.

Researchers have decided to set the margin of equivalence to 0.20. Past experience leads the researchers to set the COV to 1.50. The significance level is 0.05. The power will be computed assuming that the true ratio is one. Sample sizes between 50 and 550 will be included in the analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Higher Is	**Good**
Test Type	**Non-Inferiority**
E (Equivalence Margin)	**0.20**
R1 (True Ratio)	**1.0**
COV (Coefficient of Variation)	**1.50**
N (Total Sample Size)	**50 to 550 by 100**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*

Annotated Output

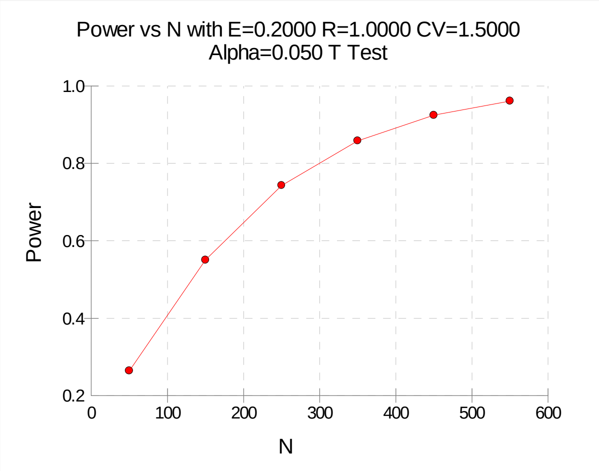
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority Ratio Test (H0: R <= 1-E; H1: R > 1-E)

                      Relative          Ratio
                   Equivalence    Equivalence     Actual   Significance
                        Margin          Bound      Ratio          Level
Power          N           (E)           (RB)       (R1)        (Alpha)       Beta        COV
0.2638        50        0.2000         0.8000     1.0000         0.0500     0.7362     1.5000
0.5505       150        0.2000         0.8000     1.0000         0.0500     0.4495     1.5000
0.7411       250        0.2000         0.8000     1.0000         0.0500     0.2589     1.5000
0.8574       350        0.2000         0.8000     1.0000         0.0500     0.1426     1.5000
0.9241       450        0.2000         0.8000     1.0000         0.0500     0.0759     1.5000
0.9607       550        0.2000         0.8000     1.0000         0.0500     0.0393     1.5000

Report Definitions
H0 (null hypothesis) is that R <= 1-E, where R = Treatment Mean / Reference Mean.
H1 (alternative hypothesis) is that R > 1-E.
E is the magnitude of the relative margin of equivalence.
RB is equivalence bound for the ratio.
R1 is actual ratio between the treatment and reference means.
COV is the coefficient of variation on the original scale.
Power is the probability of rejecting H0 when it is false.
N is the total sample size drawn from all sequences. The sample is divided equally among sequences.
Alpha is the probability of falsely rejecting H0.
Beta is the probability of not rejecting H0 when it is false.

Summary Statements
A total sample size of 50 achieves 26% power to detect non-inferiority using a one-sided t-test
when the relative margin of equivalence is 0.2000, the true mean ratio is 1.0000, the
significance level is 0.0500, and the coefficient of variation on the original, unlogged scale
is 1.5000. A 2x2 cross-over design with an equal number in each sequence is used.
```

This report shows the power for the indicated scenarios. Note that if they want 90% power, they will require a sample of around 450 subjects.

Plot Section



This plot shows the power versus the sample size.

Example 2 – Validation

We could not find a validation example for this procedure in the statistical literature. Therefore, we will show that this procedure gives the same results as the non-inferiority test on differences—a procedure that has been validated. We will use the same settings as those given in Example 1. Since the output for this example is shown above, only the output from the procedure that uses differences is shown below.

To run the inferiority test on differences, we need the values of |E| and Sw.

$$\begin{aligned} Sw &= \sqrt{\ln\left(COV^2+1\right)} \\ &= \sqrt{\ln\left(1.5^2+1\right)} \\ &= 1.085659 \end{aligned}$$

$$\begin{aligned} E &= \sqrt{\ln\left(1-E\right)} \\ &= \sqrt{\ln\left(0.8\right)} \\ &= 0.223144 \end{aligned}$$

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1a** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.223144**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**1.085659**
N (Total Sample Size)	**50 to 550 by 100**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Non-Inferiority T-Test (H0: D <= -|E|; H1: D > -|E|)

                  Equivalence     Actual  Significance                Standard
                       Margin Difference         Level               Deviation
Power          N          (E)        (D)       (Alpha)        Beta        (Sw)
0.2638        50       -0.223      0.000        0.0500      0.7362       1.086
0.5505       150       -0.223      0.000        0.0500      0.4495       1.086
0.7411       250       -0.223      0.000        0.0500      0.2589       1.086
0.8574       350       -0.223      0.000        0.0500      0.1426       1.086
0.9242       450       -0.223      0.000        0.0500      0.0758       1.086
0.9607       550       -0.223      0.000        0.0500      0.0393       1.086
```

You can compare these power values with those shown above in Example 1 to validate the procedure. You will find that the power values are identical.

Chapter 520

Equivalence Tests for Two Means in a 2x2 Cross-Over Design using Differences

Introduction

This procedure calculates power and sample size of statistical tests of equivalence of the means of a 2x2 cross-over design which is analyzed with a t-test. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. Only a brief introduction to the subject will be given here. For a comprehensive discussion on the subject, refer to Chow and Liu (1999) and Julious (2004).

Measurements are made on individuals that have been randomly assigned to one of two sequences. The first sequence receives the treatment followed by the reference (AB). The second sequence receives the reference followed by the treatment (BA). This *cross-over* design may be analyzed by a TOST equivalence test to show that the two means do not differ by more than a small amount, called the margin of equivalence.

The definition of equivalence has been refined in recent years using the concepts of prescribability and switchability. *Prescribability* refers to ability of a physician to prescribe either of two drugs at the beginning of the treatment. However, once prescribed, no other drug can be substituted for it. *Switchability* refers to the ability of a patient to switch from one drug to another during treatment without adverse effects. Prescribability is associated with equivalence of location and variability. Switchability is associated with the concept of individual equivalence. This procedure analyzes average equivalence. Thus, it partially analyzes prescribability. It does not address equivalence of variability or switchability.

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments and the objective is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to the other. It is assumed that there is a *washout* period between treatments during which the response returns back to its baseline value. If this does not occur, there is said to be a *carryover* effect.

A 2x2 cross-over design contains to two *sequences* (treatment orderings) and two time periods (occasions). One sequence receives treatment A followed by treatment B. The other sequence receives B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carry over to the second. Thus, the groups in this design are defined by the sequence in which the drugs are administered, not by the treatments they receive. Indeed, higher-order cross-over designs have been used in which the same treatment is used at both occasions.

Cross-over designs are employed because, if the no-carryover assumption is met, treatment differences are measured within a subject rather than between subjects—making a more precise measurement. Examples of the situations that might use a cross-over design are the comparison of anti-inflammatory drugs in arthritis and the comparison of hypotensive agents in essential hypertension. In both of these cases, symptoms are expected to return to their usual baseline level shortly after the treatment is stopped.

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment into the study may be easier because each patient will receive both treatments. Finally, it is often more difficult to obtain a subject than to obtain a measurement.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. It may be difficult to separate the treatment effect from the period effect, the carry-over effect of the previous treatment, and the interaction between period and treatment.

The design cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it should not be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it is often more difficult to keep patients enrolled in the study. It is arguably simpler to measure a subject once than to obtain their measurement twice. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

Outline of an Equivalence Test

PASS follows the *two one-sided tests* approach described by Schuirmann (1987) and Phillips (1990). It will be convenient to adopt the following specialize notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	|E|	*Margin of equivalence.* This is a tolerance value that defines the maximum change that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value is shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	D	*True difference*. This is the value of $\mu_T - \mu_R$, the difference between the treatment and reference means. This is the value at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their difference is needed for power and sample size calculations.

The null hypothesis of non-equivalence is

$$\mathrm{H_0}{:}\delta \le \varepsilon_L \text{ or } \delta \ge \varepsilon_U \textit{ where } \varepsilon_L < 0, \varepsilon_U > 0.$$

and the alternative hypothesis of equivalence is

$$\mathrm{H_1}{:}\varepsilon_L < \delta < \varepsilon_U$$

Test Statistics

This section describes the test statistic that is used to perform the hypothesis test.

T-Test

A t-test is used to analyze the data. The test assumes that the data are a simple random sample from a population of normally-distributed values that have the same variance. This assumption implies that the differences are continuous and normal. The calculation of the two, one-sided t-tests proceeds as follow

$$T_L = \frac{\left(\bar{x}_T - \bar{x}_R\right) - \varepsilon_L}{\hat{\sigma}_w \sqrt{\frac{2}{N}}} \text{ and } T_U = \frac{\left(\bar{x}_T - \bar{x}_R\right) - \varepsilon_U}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

where $\hat{\sigma}_w^2$ is the within mean square error from the appropriate ANOVA table.

The significance of each test statistic is determined by computing the p-value. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected.

If prior studies used a t-test rather than an ANOVA to analyze the data, you may not have a direct estimate of σ_w^2. Instead, you will have an estimate of the variance of the period differences from the t-test, $\hat{\sigma}_d^2$. These variances are functionally related by $\sigma_w^2 = 2\sigma_d^2$. Either variance can be entered.

Power Calculation

The power of this test is given by

$$\Pr(T_L \ge t_{1-\alpha,N-2} \;\text{ and }\; T_U \le -t_{1-\alpha,N-2})$$

where T_L and T_U are distributed as the bivariate, noncentral *t* distribution with noncentrality parameters Δ_L and Δ_U given by

$$\mathit{\Delta}_L = \frac{\delta - \varepsilon_L}{\sigma_w \sqrt{\dfrac{2}{N}}}$$

$$\mathit{\Delta}_U = \frac{\delta - \varepsilon_U}{\sigma_w \sqrt{\dfrac{2}{N}}}$$

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level.

Select *Power and Beta* when you want to calculate the power of an experiment that has already been run.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of unequal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of unequal means when in fact the means are unequal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in both sequences). This value must be an integer greater than one.

When N is even, it is split evenly between the two sequences. When N is odd, the first sequence has one more subject than the second sequence.

Note that you may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Equivalence Limits

|EU| (Upper Equivalence Limit)

This value gives upper limit on equivalence. Differences outside EL and EU are not considered equivalent. Differences between them are considered equivalent.

Note that EL<0 and EU>0. Also, you must have EL<D<EU.

-|EL| (Lower Equivalence Limit)

This value gives lower limit on equivalence. Differences outside EL and EU are not considered equivalent. Differences between them are.

If you want symmetric limits, enter -UPPER LIMIT for EL to force EL = -|EU|.

Note that EL<0 and EU>0. Also, you must have EL<D<EU. Finally, the scale of these numbers must match the scale of S.

Effect Size – True Mean Difference

D (True Difference)

This is the true difference between the two means at which the power is to be computed. Often this value is set to zero, but it can be non-zero as long as it is between the equivalence limits EL and EU.

Effect Size – Standard Deviation

Specify S as Sw or Sd

Specify the form of the standard deviation that is entered in the box below.

- **Sw**

  Specify S as the square root of the within mean square error from a repeated measures ANOVA. This is the most common method since cross-over designs are usually analyzed using ANOVA.

- **Sd**

  Specify S as the standard deviation of the individual treatment differences computed for each subject. This option is used when you have previous studies that produced this value.

S (Value of Sw or Sd)

Specify the value(s) of the standard deviation S. The interpretation of this value depends on the entry in *Specify S as Sw or Sd* above. If S = Sw is selected, this is the value of Sw which is SQR(WMSE) where WMSE is the within mean square error from the ANOVA table used to analyze the Cross-Over design. If S = Sd is selected, this is the value of Sd which is the standard deviation of the period differences—pooled from both sequences.

These values must be positive. A list of values may be entered.

You can press the SD button to load the Standard Deviation Estimator window.

Example 1 – Finding Power

A cross-over design is to be used to compare the impact of two drugs on diastolic blood pressure. The average diastolic blood pressure after administration of the reference drug is known to be 96 mmHg. Researchers believe this average may drop to 92 mmHg with the use of a new drug. The within mean square error of similar studies is 324. Its square root is 18.

Following FDA guidelines, the researchers want to show that the diastolic blood pressure with the new drug is within 20% of the diastolic blood pressure with the reference drug. Thus, the equivalence limits of the mean difference of the two drugs are -19.2 and 19.2. They decide to calculate the power for a range of sample sizes between 6 and 100. The significance level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**6 10 16 20 40 60 80 100**
|EU| (Upper Equivalence Limit)	**19.2**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**-4**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**18**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design

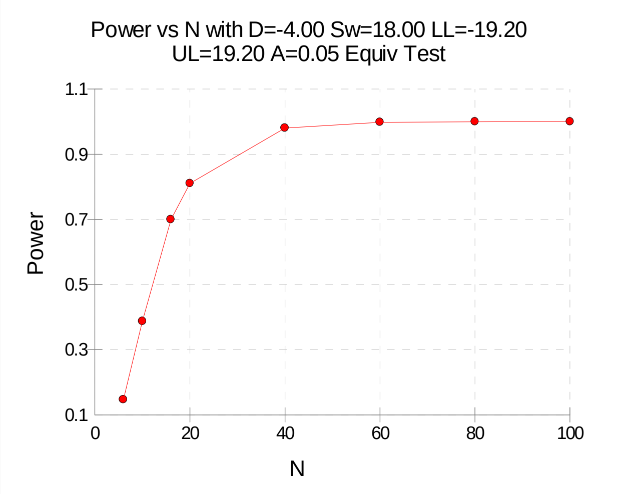
```
           Total
          Sample      Lower     Upper                 Standard
            Size     Equiv.    Equiv.        True    Deviation
Power        (N)      Limit     Limit  Difference           Sw     Alpha      Beta
0.1470         6     -19.20     19.20       -4.00        18.00    0.0500    0.8530
0.3873        10     -19.20     19.20       -4.00        18.00    0.0500    0.6127
0.6997        16     -19.20     19.20       -4.00        18.00    0.0500    0.3003
0.8104        20     -19.20     19.20       -4.00        18.00    0.0500    0.1896
0.9804        40     -19.20     19.20       -4.00        18.00    0.0500    0.0196
0.9983        60     -19.20     19.20       -4.00        18.00    0.0500    0.0017
0.9999        80     -19.20     19.20       -4.00        18.00    0.0500    0.0001
1.0000       100     -19.20     19.20       -4.00        18.00    0.0500    0.0000
```

Report Definitions
Power is the probability of rejecting non-equivalence when the means are equivalent.
N is the total number of subjects split between both sequences.
EU & EL are the maximum allowable differences that still result in equivalence.
D is the difference between the means at which the power is computed.
Sw is the square root of the within mean square error from the ANOVA table.
Alpha is the probability of rejecting non-equivalence when the means are non-equivalent.
Beta is the probability of accepting non-equivalence when the means are equivalent.

Summary Statements
In an equivalence test of means using two one-sided tests on data from a two-period cross-over design, a total sample size of 6 achieves 15% power at a 5% significance level when the true difference between the means is -4.00, the square root of the within mean square error is 18.00, and the equivalence limits are -19.20 and 19.20.

This report shows the power for the indicated scenarios. Note that if they want 90% power, they will require a sample of around 30 subjects.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Finding Sample Size

Continuing with Example 1, the researchers want to find the exact sample size needed to achieve both 80% power and 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.8 0.9**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
|EU| (Upper Equivalence Limit)	**19.2**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**-4**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**18**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Cross-Over Design

              Total
             Sample     Lower     Upper                 Standard
               Size    Equiv.    Equiv.       True     Deviation
Power           (N)     Limit     Limit   Difference          Sw     Alpha       Beta
0.9032           26    -19.20     19.20       -4.00        18.00    0.0500     0.0968
0.8104           20    -19.20     19.20       -4.00        18.00    0.0500     0.1896
```

We note that 20 subjects are needed to achieve 80% power and 26 subjects are needed to achieve 90% power.

Example 3 – Validation using Phillips

Phillips (1990) page 142 presents a table of sample sizes for various parameter values. In this table, the treatment mean, standard deviation, and equivalence limits are all specified as percentages of the reference mean. We will reproduce the second line of the table in which the square root of the within mean square error is 20%; the equivalence limits are 20%; the treatment mean is 100%, 95%, 90%, and 85%; the power is 70%; and the significance level is 0.05. Phillips reports total sample size as 16, 20, 40, and 152 corresponding to the four treatment mean percentages. We will now setup this example in ***PASS***.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.7**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
|EU| (Upper Equivalence Limit)	**20**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**0 -5 -10 -15**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**20**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Cross-Over Design

           Total
          Sample    Lower    Upper                Standard
            Size   Equiv.   Equiv.        True   Deviation
Power        (N)    Limit    Limit  Difference          Sw    Alpha     Beta
0.7001       152   -20.00    20.00      -15.00       20.00   0.0500   0.2999
0.7092        40   -20.00    20.00      -10.00       20.00   0.0500   0.2908
0.7221        20   -20.00    20.00       -5.00       20.00   0.0500   0.2779
0.7031        16   -20.00    20.00        0.00       20.00   0.0500   0.2969
```

Note that ***PASS*** has obtained the same samples sizes as Phillips (1990).

Example 4 – Validation using Machin

Machin *et al.* (1997) page 107 present an example of determining the sample size for a cross-over design in which the reference mean is 35.03, the treatment mean is 35.03, the standard deviation, entered as the square root of the within mean square error, is 40% of the reference mean, the limits are plus or minus 20% of the reference mean, the power is 80%, and the significance level is 0.10. Machin *et al.* calculate the total sample size to be 54.

When the parameters are given as percentages of the reference mean, it is easy enough to calculate the exact amounts by applying those percentages. However, the percentages can all be entered directly as long as all parameters (EU, EL, D, and Sw) are specified as percentages.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.8**
Alpha	**0.10**
N (Total Sample Size)	*Ignored since this is the Find setting*
|EU| (Upper Equivalence Limit)	**20**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**40**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Cross-Over Design

            Total
           Sample    Lower    Upper                 Standard
             Size   Equiv.   Equiv.        True    Deviation
Power         (N)    Limit    Limit  Difference           Sw     Alpha      Beta
0.8050         54   -20.00    20.00        0.00        40.00    0.1000    0.1950
```

Note that ***PASS*** also has obtained a sample size of 54.

Example 5 – Validation using Chow and Liu

Chow and Liu (1999) page 153 present an example of determining the sample size for a cross-over design in which the reference mean is 82.559, the treatment mean is 82.559, the standard deviation, entered as the square root of the within mean square error, is 15.66%, the limits are plus or minus 20%, the power is 80%, and the significance level is 0.05. They calculate a sample size of 12. ***PASS*** calculates a sample size of 13. To see why ***PASS*** has increased the sample size by one, we will evaluate the power at sample sizes of 10, 12, 13, 14, and 16.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**10 12 13 14 16**
|EU| (Upper Equivalence Limit)	**20**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**15.66**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Cross-Over Design

             Total
            Sample     Lower    Upper                Standard
              Size    Equiv.   Equiv.        True   Deviation
Power          (N)     Limit    Limit  Difference          Sw     Alpha      Beta
0.6643          10    -20.00    20.00        0.00       15.66    0.0500    0.3357
0.7932          12    -20.00    20.00        0.00       15.66    0.0500    0.2068
0.8363          13    -20.00    20.00        0.00       15.66    0.0500    0.1637
0.8752          14    -20.00    20.00        0.00       15.66    0.0500    0.1248
0.9258          16    -20.00    20.00        0.00       15.66    0.0500    0.0742
```

The power for N = 12 is 0.7932. The power for N = 13 is 0.8363. Hence, to achieve better than 80% power, a sample size of 13 is necessary. However, 0.7932 is sufficiently close to 0.800 to make N = 12 a reasonable choice (as Chow and Liu did).

Example 6 – Validation using Senn

Senn (1993) page 217 presents an example of determining the sample size for a cross-over design in which the reference mean is equal to the treatment mean, the standard deviation, entered as the square root of the within mean square error, is 45, the equivalence limits are plus or minus 30, the power is 80%, and the significance level is 0.05. He calculates a sample size of 40.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.8**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
|EU| (Upper Equivalence Limit)	**30**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**0**
Specify S as Sw or Sd	**Sw**
S (Value of Sw or Sd)	**45**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Cross-Over Design

              Total
             Sample    Lower    Upper              Standard
               Size   Equiv.   Equiv.       True  Deviation
Power           (N)    Limit    Limit Difference         Sw     Alpha      Beta
0.8004           40   -30.00    30.00       0.00      45.00    0.0500    0.1996
```

PASS also calculates a sample size of 40.

Chapter 525

Equivalence Tests for Two Means in a 2x2 Cross-Over Design using Ratios

Introduction

This procedure calculates power and sample size of statistical tests of equivalence of the means from a 2x2 cross-over design which is analyzed with a t-test. This routine deals with the case in which the statistical hypotheses are expressed in terms mean of ratios rather than mean differences.

The details of testing the equivalence of two treatments using data from a 2x2 cross-over design are given in another chapter and will not be repeated here. If the logarithms of the responses can be assumed to follow the normal distribution, hypotheses about the equivalence of two means stated in terms of the ratio can be transformed into hypotheses about the difference. The details of this analysis are given in Julious (2004). They will only be summarized here.

Equivalence Testing Using Ratios

PASS follows the *two one-sided tests* approach described by Schuirmann (1987) and Phillips (1990). It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ϕ_L, ϕ_U	RL, RU	*Margin of equivalence*. These limits that define an interval of the ratio of the means in which their difference is so small that it may be ignored.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their ratio is needed for power and sample size calculations.

The null hypothesis of non-equivalence is

$$\mathrm{H}_0{:}\phi \le \phi_L \text{ or } \phi \ge \phi_U \quad where\ \phi_L < 1,\ \phi_U > 1.$$

and the alternative hypothesis of equivalence is

$$\mathrm{H}_1{:}\phi_L < \phi < \phi_U$$

Log-Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as scale-less percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of ratios.
2. Transform these into hypotheses about differences by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the null hypothesis are as follows.

$$\begin{aligned}
&\phi_L \le \phi \le \phi_U \\
&\Rightarrow \phi_L \le \left\{\frac{\mu_T}{\mu_R}\right\} \le \phi_U \\
&\Rightarrow \ln(\phi_L) \le \{\ln(\mu_T) - \ln(\mu_R)\} \le \ln(\phi_U)
\end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

When performing an equivalence test on the difference between means, the usual procedure is to set the equivalence limits symmetrically above and below zero. Thus the equivalence limits will be plus or minus an appropriate amount. The common practice is to do the same when the data are being analyzed on the log scale. However, when symmetric limits are set on the log scale, they do not translate to symmetric limits on the original scale. Instead, they translate to limits that are the inverses of each other.

Perhaps these concepts can best be understood by considering an example. Suppose the researchers have determined that the lower equivalence limit should be 80% on the original scale. Since they are planning to use a log scale for their analysis, they transform this limit to the log scale by taking the logarithm of 0.80. The result is -0.223144. Wanting symmetric limits, they set the upper equivalence limit to 0.223144. Exponentiating this value, they find that exp(0.223144) = 1.25. Note that 1/(0.80) = 1.25. Thus, the limits on the original scale are 80% and 125%, not 80% and 120%.

Using this procedure, appropriate equivalence limits for the ratio of two means can be easily determined. Here are a few sets of equivalence limits.

Specified Percent Change	Lower Limit Original Scale	Upper Limit Original Scale	Lower Limit Log Scale	Upper Limit Log Scale
-25%	75.0%	133.3%	-0.287682	0.287682
+25%	80.0%	125.0%	-0.223144	0.223144
-20%	80.0%	125.0%	-0.223144	0.223144
+20%	83.3%	120.0%	-0.182322	0.182322
-10%	90.0%	111.1%	-0.105361	0.105361
+10%	90.9%	110.0%	-0.095310	0.095310

Note that negative percent-change values specify the lower limit first, while positive percent-change values specify the upper limit first. After the first limit is found, the other limit is calculated as its inverse.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter can be used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable X is the logarithm of the original variable Y. That is, X = ln(Y) and Y = exp(X). Label the mean and variance of X as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of Y as μ_Y and σ_Y^2, respectively. If X is normally distributed, then Y is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of Y can be expressed as

$$
\begin{aligned}
COV_Y &= \frac{\sqrt{\mu_Y^2\left(e^{\sigma_X^2}-1\right)}}{\mu_Y} \\
&= \sqrt{e^{\sigma_X^2}-1}
\end{aligned}
$$

Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2+1\right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2+1\right)}$$

Thus, the hypotheses can be stated in the original (Y) scale and then the power can be analyzed in the transformed (X) scale.

Power Calculation

As is shown above, the hypotheses can be stated in the original (Y) scale using ratios or the logged (X) scale using differences. Either way, the power and sample size calculations are made using the formulas for testing the equivalence of the difference in two means. These formulas are presented in another chapter and are not duplicated here.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of unequal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of unequal means when in fact the means are unequal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in both sequences). This value must be an integer greater than one.

When N is even, it is split evenly between the two sequences. When N is odd, the first sequence has one more subject than the second sequence.

Note that you may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Equivalence Limits

RU (Upper Equivalence Limit)

Enter the upper equivalence limit for the ratio of the two means. When the ratio of the means is between this value and RL, the two means are said to be equivalent. The value must be greater than one. A popular choice is 1.25. Note that this value is not a percentage.

If you enter *1/RL*, then 1/RL will be calculated and used here. This choice is commonly used because RL and 1/RL give limits that are of equal magnitude on the log scale.

RL (Lower Equivalence Limit)

Enter the lower equivalence limit for the ratio of the two means. When the ratio of the means is between this value and RU, the two means are said to be equivalent. The value must be less than one. A popular choice is 0.80. Note that this value is not a percentage.

If you enter *1/RU*, then 1/RU will be calculated and used here. This choice is commonly used because RU and 1/RU give limits that are of equal magnitude on the log scale.

Effect Size – True Ratio

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Usually, the ratio will be assumed to be one. However, some authors recommend calculating the power using a ratio of 1.05 since this will require a larger sample size.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not logged) scale. This value must be determined from past experience or from a pilot study. It is most easily calculated from the within mean-square error of the analysis of variance using the logged data using the relationship

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\,.$$

If prior studies used a t-test to analyze the logged data, you will not have a direct estimate of $\hat{\sigma}_w^2$. However, the two variances, σ_d^2 and σ_w^2, are functionally related. The relationship between these quantities is $\sigma_d^2 = 2\sigma_w^2$.

Example 1 – Finding Power

A company has opened a new manufacturing plant and wants to show that the drug produced in the new plant is equivalent to that produced in an older plant. A cross-over design will be used to test the equivalence of drugs produced at the two plants.

Researchers have decided to set the equivalence limits for the ratio at 0.90 and 1.111 (note that 1.111 = 1/0.90). Past experience leads the researchers to set the COV to 0.50. The significance level is 0.05. The power will be computed assuming that the true ratio is one. Sample sizes between 50 and 550 will be included in the analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**50 to 550 by 100**
RU (Upper Equivalence Limit)	**1/RL**
RL (Lower Equivalence Limit)	**0.90**
R1 (True Ratio)	**1.0**
COV (Coefficient of Variation)	**0.50**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing Equivalence Using a Cross-Over Design

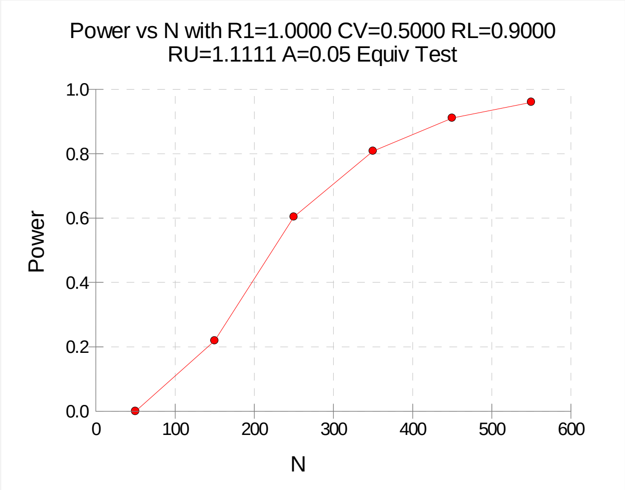
```
                  Lower      Upper
         Total    Equiv.     Equiv.              Coefficient
        Sample     Limit      Limit       True            of
          Size  of Ratio   of Ratio      Ratio     Variation
Power      (N)      (RL)       (RU)       (R1)         (COV)     Alpha       Beta
0.0000      50    0.9000     1.1111     1.0000        0.5000    0.0500     1.0000
0.2190     150    0.9000     1.1111     1.0000        0.5000    0.0500     0.7810
0.6002     250    0.9000     1.1111     1.0000        0.5000    0.0500     0.3998
0.8064     350    0.9000     1.1111     1.0000        0.5000    0.0500     0.1936
0.9101     450    0.9000     1.1111     1.0000        0.5000    0.0500     0.0899
0.9596     550    0.9000     1.1111     1.0000        0.5000    0.0500     0.0404
```

Report Definitions
Power is the probability of rejecting non-equivalence when the means are equivalent.
N is the total number of subjects split between both sequences.
RU & RL are the upper and lower equivalence limits. Ratios between these limits are equivalent.
R1 is the ratio of the means at which the power is computed.
COV is the coefficient of variation on the original scale.
Alpha is the probability of rejecting non-equivalence when the means are non-equivalent.
Beta is the probability of accepting non-equivalence when the means are equivalent.

Summary Statements
In an equivalence test of means using two one-sided tests on data from a two-period cross-over design, a total sample size of 50 achieves 0% power at a 5% significance level when the true ratio of the means is 1.0000, the coefficient of variation on the original, unlogged scale is 0.5000, and the equivalence limits of the mean ratio are 0.9000 and 1.1111.

This report shows the power for the indicated scenarios. Note that if they want 90% power, they will require a sample of around 450 subjects.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Validation using Julious

Julious (2004) page 1963 presents a table of sample sizes for various parameter values. The power is 0.90 and the significance level is 0.05. The COV is set to 0.25, the ‘level of bioequivalence’ is set to 10%, 15%, 20%, and 25%, and the true ratio is set to 1.00, the necessary sample sizes are 120, 52, 28, and 18. Note that the level of bioequivalence as defined in Julious (2004) is equal to 1 – RL.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a 2x2 Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **2x2 Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N**
Power	**0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
RU (Upper Equivalence Limit)	**1/RL**
RL (Lower Equivalence Limit)	**0.90 0.85 0.80 0.75**
R1 (True Ratio)	**1.00**
COV (Coefficient of Variation)	**0.25**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Equivalence Using a Cross-Over Design

                      Lower     Upper
              Total   Equiv.    Equiv.             Coefficient
             Sample   Limit     Limit     True          of
               Size   of Ratio  of Ratio  Ratio    Variation
Power           (N)     (RL)      (RU)     (R1)        (COV)     Alpha      Beta
0.9121           18   0.7500    1.3333   1.0000       0.2500    0.0500    0.0879
0.9023           28   0.8000    1.2500   1.0000       0.2500    0.0500    0.0977
0.9060           52   0.8500    1.1765   1.0000       0.2500    0.0500    0.0940
0.9012          120   0.9000    1.1111   1.0000       0.2500    0.0500    0.0988
```

Note that ***PASS*** obtains the same samples sizes as Julious (2004).

Chapter 530

Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design using Differences

Introduction

This procedure calculates power and sample size for non-inferiority and superiority tests which use the difference in the means of a higher-order cross-over design. Measurements are made on individuals that have been randomly assigned to one of several treatment sequences. Only a brief introduction to the subject will be given here. For a comprehensive discussion on the subject, refer to Chen et al. (1997) and Chow et al. (2003).

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments at least once and the object is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to another. It is assumed that there is a *washout* period

between treatments during which the response returns to its baseline value. If this does not occur, there is said to be a *carryover* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence of treatments is treatment A followed by treatment B. The other sequence is B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carryover to the second. Thus, the groups of subjects in this design are defined by the sequence in which the two treatments are administered, not by the treatments they receive.

Higher-Order Cross-Over Designs

Chen et al. (1997) present the results for four cross-over designs that are more complicated than the 2x2 design. Assume that the two treatments are labeled A and B. The available designs are defined by the order and number of times the two treatments are administered.

Balaam's Design

Balaam's design has four sequences with two treatments each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2
1	A	A
2	B	B
3	A	B
4	B	A

Two-Sequence Dual Design

This design has two sequences with three periods each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2	Period 3
1	A	B	B
2	B	A	A

Four-Period Design with Two Sequences

This design has two sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	B	B	A
2	B	A	A	B

Four-Period Design with Four Sequences

This design has four sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	A	B	B
2	B	B	A	A
1	A	B	B	A
2	B	A	A	B

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment may be easier to obtain because each patient will receive both treatments.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. In a cross-over experiment, it may be difficult to separate the treatment effect from the time effect and the carry-over effect of the previous treatment.

These cross-over designs cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it cannot be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it may be more difficult to keep patients enrolled in the study. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests. Remember that in the usual t-test setting, the null (H0) and alternative (H1) hypotheses for one-sided tests are defined as

$$\mathrm{H}_0 : \delta \le A \text{ versus } \mathrm{H}_1 : \delta > A$$

Rejecting H0 implies that the mean is larger than the value *A*. This test is called an *upper-tailed test* because H0 is rejected only in samples in which the difference in sample means is larger than *A*.

Following is an example of a *lower-tailed test*.

$$\mathrm{H}_0 : \delta \ge A \text{ versus } \mathrm{H}_1 : \delta < A$$

Non-inferiority and *superiority* tests are special cases of the above directional tests. It will be convenient to adopt the following specialize notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	|E|	*Margin of equivalence.* This is a tolerance value that defines the maximum difference that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value symbols are shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	D	*True difference*. This is the value of $\mu_T - \mu_R$, the difference between the treatment and reference means. This is the value at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their difference is needed for power and sample size calculations.

Non-Inferiority and Superiority Tests

A *non-inferiority test* tests that the treatment mean is not worse than the reference mean by more than a small equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

A *superiority test* tests that the treatment mean is better than the reference mean by more than a small equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

Case 1: High Values Good, Non-Inferiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no less than a small amount below the reference mean. The value of δ is often set to zero. The null and alternative hypotheses are

$$\mathrm{H}_0: \mu_T \le \mu_R - |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1: \mu_T > \mu_R - |\varepsilon|$$

$$\mathrm{H}_0: \mu_T - \mu_R \le -|\varepsilon| \quad \text{versus} \quad \mathrm{H}_1: \mu_T - \mu_R > -|\varepsilon|$$

$$\mathrm{H}_0: \delta \le -|\varepsilon| \quad \text{versus} \quad \mathrm{H}_1: \delta > -|\varepsilon|$$

Case 2: High Values Bad, Non-Inferiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no more than a small amount above the reference mean. The value of δ is often set to zero. The null and alternative hypotheses are

$$\mathrm{H}_0{:}\mu_T \geq \mu_R + |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\mu_T < \mu_R + |\varepsilon|$$

$$\mathrm{H}_0{:}\mu_T - \mu_R \geq |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\mu_T - \mu_R < |\varepsilon|$$

$$\mathrm{H}_0{:}\delta \geq |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\delta < |\varepsilon|$$

Case 3: High Values Good, Superiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is greater than the reference mean by at least the margin of equivalence. The specified value of δ must be greater than the specified value of $|\varepsilon|$. The null and alternative hypotheses are

$$\mathrm{H}_0{:}\mu_T \leq \mu_R + |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\mu_T > \mu_R + |\varepsilon|$$

$$\mathrm{H}_0{:}\mu_T - \mu_R \leq |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\mu_T - \mu_R > |\varepsilon|$$

$$\mathrm{H}_0{:}\delta \leq |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\delta > |\varepsilon|$$

Case 4: High Values Bad, Superiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is less than the reference mean by at least the margin of equivalence. The specified value of δ must be less than the specified value of $-|\varepsilon|$. The null and alternative hypotheses are

$$\mathrm{H}_0{:}\mu_T \geq \mu_R - |\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\mu_T < \mu_R - |\varepsilon|$$

$$\mathrm{H}_0{:}\mu_T - \mu_R \geq -|\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\mu_T - \mu_R < -|\varepsilon|$$

$$\mathrm{H}_0{:}\delta \geq -|\varepsilon| \quad \text{versus} \quad \mathrm{H}_1{:}\delta < -|\varepsilon|$$

Test Statistics

The analysis for assessing equivalence (and thus non-inferiority) using higher-order cross-over designs is discussed in detail in Chapter 9 of Chow and Liu (2000). Unfortunately, their presentation is too lengthy to give here. Their method involves the computation of an analysis of variance to estimate the error variance. It also describes the construction of confidence limits for appropriate contrasts. One-sided confidence limits can be used for non-inferiority tests. Details of this approach are given in Chapter 3 of Chow et al. (2003). We refer you to these books for details.

Power Calculation

The power of the non-inferiority and superiority tests for the case in which higher values are better is given by

$$Power = T_V\left(\left(\frac{\delta-\varepsilon}{\sigma_W\sqrt{b/n}}\right)-t_{V,1-\alpha}\right)$$

where T represents the cumulative t distribution, V and b depend on the design, σ_W is the square root of the within mean square error from the ANOVA table used to analyze the cross-over design, and n is the average number of subjects per sequence. Note that the constants V and b depend on the design as follows.

The power of the non-inferiority and superiority tests for the case in which higher values are worse is given by

$$Power = 1-T_V\left(t_{V,1-\alpha}-\left(\frac{\varepsilon-\delta}{\sigma_W\sqrt{b/n}}\right)\right)$$

The constants V and b depend on the design as follows:

Design Type	**Parameters (*V*,*b*)**
Balaam’s Design	V = 4n – 3, b = 2.
Two-Sequence Dual Design	V = 4n – 4, b = 3/4.
Four-Period Design with Two Sequences	V = 6n – 5, b = 11/20.
Four-Period Design with Four Sequences	V = 12n – 5, b = 1/4.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level. Note that there are two choices for finding *N*. Select *N (Equal Per Sequence)* when you want the design to have an equal number of subjects per sequence. Select *N (Exact)* when you want to find the exact sample size even though the number of subjects cannot be dividing equally among the sequences.

Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in all sequences). This value must be an integer greater than one.

Effect Size – Mean Difference

|E| (Equivalence Margin)

This is the magnitude of the *margin of equivalence*. It is the smallest difference between the treatment and reference means that still results in the conclusion of non-inferiority (or superiority). The sign of this value is assigned depending on the selections for Higher Is and Test Type.

D (True Difference)

This is the true difference between the two means at which the power is to be computed. Often this value is set to zero, but it can be non-zero as long as it is between zero and |E|.

Effect Size – Standard Deviation

Sw (Within Standard Error)

Specify one or more values of Sw, which is SQR(WMSE) where WMSE is the within mean square error from the ANOVA table used to analyze the cross-over design. These values must be positive.

You can press the Standard Deviation Estimator button to load the Standard Deviation Estimator window.

Test

Design Type

Specify the type of cross-over design that you are analyzing. Note that all of these designs assume that you are primarily interested in the overall difference between the two treatment means.

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the treatment mean is within the margin of equivalence of being no worse than the reference mean. Select *Superiority* when you want to test whether the treatment mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are often considered bad. However, if the response variable is income, higher values are usually considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Example 1 – Finding Power

Researchers want to calculate the power of a non-inferiority test using data from a two-sequence, dual cross-over design. The margin of equivalence is either 5 or 10 at several sample sizes between 6 and 66. The true difference between the means under is assumed to be 0. Similar experiments have had a standard deviation (Sw) of 10. The significance level is 0.025.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.025**
N (Total Sample Size)	**6 to 66 by 10**
|E| (Equivalence Margin)	**5 10**
D (True Difference)	**0**
Sw (Within Standard Error)	**10**
Design Type	**2x3 (Two-Sequence Dual)**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Non-Inferiority Using the Difference
Design: Two-Sequence Dual Cross-Over. Hypotheses: H0: D <= -|E|; H1: D > -|E|.

               Total   Sequences                 Difference   Standard
              Sample         and    Equivalence         for      Error
                Size     Periods         Margin       Power   of Diff.
Power            (N)       (SxP)            |E|         (D)       (Sw)     Alpha      Beta
0.1139             6         2x3           5.00        0.00      10.00    0.0250    0.8861
0.3405            16         2x3           5.00        0.00      10.00    0.0250    0.6595
0.5282            26         2x3           5.00        0.00      10.00    0.0250    0.4718
0.6744            36         2x3           5.00        0.00      10.00    0.0250    0.3256
0.7817            46         2x3           5.00        0.00      10.00    0.0250    0.2183
0.8571            56         2x3           5.00        0.00      10.00    0.0250    0.1429
0.9084            66         2x3           5.00        0.00      10.00    0.0250    0.0916
Report continues…
```

References

Chow, S.C. and Liu, J.P. 1999. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker. New York

Chow, S.C.; Shao, J.; Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.

Chen, K.W.; Chow, S.C.; and Li, G. 1997. 'A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs.' Journal of Pharmacokinetics and Biopharmaceutics, Volume 25, No. 6, pages 753-765.

Report Definitions

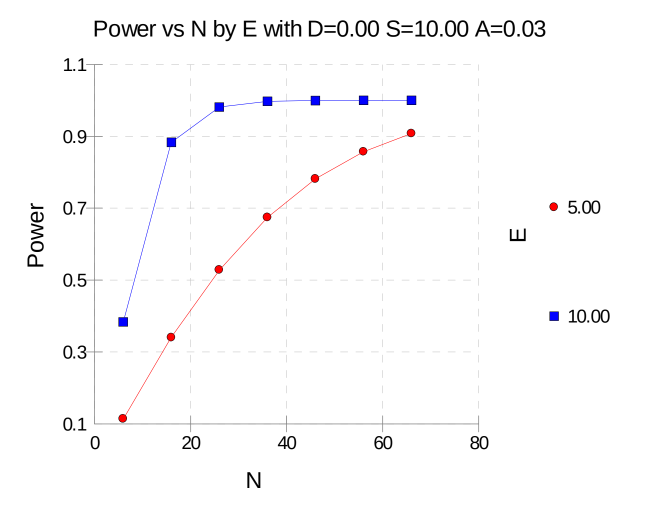
Power is the probability of rejecting H0 (concluding non-inferiority) when H0 is false.
N is the total number of subjects. They are divided evenly among all sequences.
S is the number of sequences.
P is the number of periods per sequence.
|E| is the magnitude of the margin of equivalence. It is the largest difference that is not of practical significance.
D is the difference between the means at which the power is computed.
Sw is the square root of the within mean square error from the ANOVA table.
Alpha is the probability of falsely rejecting H0 (falsely concluding non-inferiority).
Beta is the probability of not rejecting H0 when it is false.
Two-Sequence Dual Cross-Over Design with pattern: ABB; BAA

Summary Statements

In a non-inferiority test on data for which higher values are better drawn from a two-sequence dual cross-over design, a total sample size of 6 achieves 11% power at a 3% significance level when the true difference between the means is 0.00, the square root of the within mean square error is 10.00, and the equivalence margin is 5.00.

This report shows the power for the indicated scenarios.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Finding Sample Size

Continuing with Example1, the researchers want to find the exact sample size needed to achieve both 80% power and 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.80 0.90**
Alpha	**0.025**
N (Total Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**5 10**
D (True Difference)	**0**
Sw (Within Standard Error)	**10**
Design Type	**2x3 (Two-Sequence Dual)**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Non-Inferiority Using the Difference
Design: Two-Sequence Dual Cross-Over. Hypotheses: H0: D <= -|E|; H1: D > -|E|.

            Total   Sequences                Difference  Standard
           Sample         and  Equivalence        for     Error
             Size     Periods       Margin      Power   of Diff.
Power         (N)       (SxP)          |E|        (D)      (Sw)    Alpha    Beta
0.9084         66         2x3         5.00       0.00     10.00   0.0250  0.0916
0.8153         50         2x3         5.00       0.00     10.00   0.0250  0.1847
0.9184         18         2x3        10.00       0.00     10.00   0.0250  0.0816
0.8343         14         2x3        10.00       0.00     10.00   0.0250  0.1657
```

When the equivalence margin is set to 5, 66 subjects are needed to achieve 90% power and 50 subjects are needed to achieve at least 80% power.

Example 3 – Validation

We could not find a validation example for this procedure in the statistical literature, so we will have to generate a validated example from within ***PASS***. To do this, we use the Higher-Order, Cross-Over Equivalence using Differences procedure which was validated. By setting the upper equivalence limit to a large value (we used 22), we obtain results for a non-inferiority test.

Suppose the square root of the within mean square error is 0.10, the equivalence limit is 0.20, the difference between the means is 0.05, the power is 90%, and the significance level is 0.05 (see the Example4 template). ***PASS*** calculates a sample size of 16.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
|E| (Equivalence Margin)	**0.2**
D (True Difference)	**0.05**
Sw (Within Standard Error)	**0.10**
Design Type	**4x2 (Balaam)**
Data Tab (continued)	
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Non-Inferiority Using the Difference
Design: Balaam's Cross-Over. Hypotheses: H0: D <= -|E|; H1: D > -|E|.

            Total   Sequences               Difference  Standard
           Sample         and  Equivalence         for     Error
             Size     Periods       Margin       Power  of Diff.
Power         (N)       (SxP)          |E|         (D)      (Sw)     Alpha      Beta
0.9495         16         4x2         0.20        0.05      0.10    0.0500    0.0505
```

PASS has also obtained a sample size of 16 using the non-inferiority procedure.

Chapter 535

Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design using Ratios

Introduction

This procedure calculates power and sample size for non-inferiority and superiority tests which use the ratio of the two means of a higher-order cross-over design. Measurements are made on individuals that have been randomly assigned to one of several treatment sequences. Only a brief introduction to the subject will be given here. For a comprehensive discussion on the subject, refer to Chen et al. (1997) and Chow et al. (2003). - -

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments at least once and the object is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to another. It is assumed that there is a *washout* period

between treatments during which the response returns to its baseline value. If this does not occur, there is said to be a *carryover* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence of treatments is treatment A followed by treatment B. The other sequence is B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carryover to the second. Thus, the groups of subjects in this design are defined by the sequence in which the two treatments are administered, not by the treatments they receive.

Higher-Order Cross-Over Designs

Chen et al. (1997) present the results for four cross-over designs that are more complicated than the 2x2 design. Assume that the two treatments are labeled A and B. The available designs are defined by the order and number of times the two treatments are administered.

Balaam's Design

Balaam's design has four sequences with two treatments each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2
1	A	A
2	B	B
3	A	B
4	B	A

Two-Sequence Dual Design

This design has two sequences with three periods each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2	Period 3
1	A	B	B
2	B	A	A

Four-Period Design with Two Sequences

This design has two sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	B	B	A
2	B	A	A	B

Four-Period Design with Four Sequences

This design has four sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	A	B	B
2	B	B	A	A
1	A	B	B	A
2	B	A	A	B

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment may be easier to obtain because each patient will receive both treatments.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. In a cross-over experiment, it may be difficult to separate the treatment effect from the time effect and the carry-over effect of the previous treatment.

These cross-over designs cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it cannot be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it may be more difficult to keep patients enrolled in the study. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

The Statistical Hypotheses

Both non-inferiority and superiority tests are examples of directional (one-sided) tests. Remember that in the usual t-test setting, the null (H0) and alternative (H1) hypotheses for one-sided tests are defined as

$$\mathrm{H}_0\!:\phi \leq A \ \text{ versus } \ \mathrm{H}_1\!:\phi > A$$

Rejecting H0 implies that the ratio of the mean is larger than the value *A*. This test is called an *upper-tailed test* because H0 is rejected only in samples in which the ratio of the sample means is larger than *A*.

Following is an example of a *lower-tailed test*.

$$\mathrm{H}_0\!:\phi \geq A \ \text{ versus } \ \mathrm{H}_1\!:\phi < A$$

Non-inferiority and *superiority* tests are special cases of the above directional tests. It will be convenient to adopt the following specialize notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	E	*Margin of equivalence.* This is a tolerance value that defines the maximum deviation from unity that the mean ratio can be and still not be of practical importance. This is the largest change in the mean ratio from the baseline value (usually one) that is still considered to be trivial.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their ratio is needed for power and sample size calculations.

The null hypothesis of inferiority is

$$\mathrm{H}_0: \phi \le 1-\varepsilon \text{ where } \varepsilon > 0.$$

The alternative hypothesis of non-inferiority is

$$\mathrm{H}_1: \phi > 1-\varepsilon$$

Log-Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as scale-less percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypotheses in terms of ratios.
2. Transform these into hypotheses about differences by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the alternative hypothesis are as follows.

$$\begin{aligned} &1-\varepsilon < \phi \\ &\Rightarrow 1-\varepsilon < \left\{\frac{\mu_T}{\mu_R}\right\} \\ &\Rightarrow \ln(1-\varepsilon) < \left\{\ln(\mu_T) - \ln(\mu_R)\right\} \end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter is used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable *X* is the logarithm of the original variable *Y*. That is, *X* = ln(*Y*) and *Y* = exp(*X*). Label the mean and variance of *X* as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of *Y* as μ_Y and σ_Y^2, respectively. If *X* is normally distributed, then *Y* is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be found to be

$$\begin{aligned} COV_Y &= \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y} \\ &= \sqrt{e^{\sigma_X^2} - 1} \\ &= \sqrt{e^{\sigma_w^2} - 1} \end{aligned}$$

where σ_w^2 is the within mean square error from the analysis of variance of the logged data. Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2 + 1 \right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

Thus, the hypotheses can be stated in the original (Y) scale and then power analyzed in the transformed (X) scale.

Non-Inferiority and Superiority Tests

A *non-inferiority test* tests that the treatment mean is not worse than the reference mean by more than a small equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

A *superiority test* tests that the treatment mean is better than the reference mean by more than a small equivalence margin. The actual direction of the hypothesis depends on the response variable being studied.

Case 1: High Values Good, Non-Inferiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no less than a small amount below the reference mean. The null and alternative hypotheses are

$$\mathrm{H}_0\!:\frac{\mu_T}{\mu_R} \le (1-\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\frac{\mu_T}{\mu_R} > (1-\varepsilon)$$

$$\mathrm{H}_0\!:\ln(\mu_T)-\ln(\mu_R) \le \ln(1-\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\ln(\mu_T)-\ln(\mu_R) > \ln(1-\varepsilon)$$

Case 2: High Values Bad, Non-Inferiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is no more than a small amount above the reference mean. The null and alternative hypotheses are

$$\mathrm{H}_0\!:\frac{\mu_T}{\mu_R} \ge (1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\frac{\mu_T}{\mu_R} < (1+\varepsilon)$$

$$\mathrm{H}_0\!:\ln(\mu_T)-\ln(\mu_R) \ge \ln(1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\ln(\mu_T)-\ln(\mu_R) < \ln(1+\varepsilon)$$

Case 3: High Values Good, Superiority Test

In this case, higher values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is greater than the reference mean by at least the margin of equivalence. The null and alternative hypotheses are

$$\mathrm{H}_0\!:\frac{\mu_T}{\mu_R} \le (1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\frac{\mu_T}{\mu_R} > (1+\varepsilon)$$

$$\mathrm{H}_0\!:\ln(\mu_T)-\ln(\mu_R) \le \ln(1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\ln(\mu_T)-\ln(\mu_R) > \ln(1+\varepsilon)$$

Case 4: High Values Bad, Superiority Test

In this case, lower values are better. The hypotheses are arranged so that rejecting the null hypothesis implies that the treatment mean is less than the reference mean by at least the margin of equivalence. The null and alternative hypotheses are

$$\mathrm{H}_0\!:\frac{\mu_T}{\mu_R} \ge (1-\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\frac{\mu_T}{\mu_R} < (1-\varepsilon)$$

$$\mathrm{H}_0\!:\ln(\mu_T)-\ln(\mu_R) \ge \ln(1-\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1\!:\ln(\mu_T)-\ln(\mu_R) < \ln(1-\varepsilon)$$

Test Statistics

The analysis for assessing non-inferiority using higher-order cross-over designs is discussed in detail in Chapter 9 of Chow and Liu (2000). Unfortunately, their presentation is too lengthy to give here. Their method involves the computation of an analysis of variance to estimate the error variance. It also describes the construction of confidence limits for appropriate contrasts. One-sided confidence limits can be used for non-inferiority tests. Details of this approach are given in Chapter 3 of Chow et al. (2003). We refer you to these books for details.

Power Calculation

The power of the non-inferiority and superiority tests for the case in which higher values are better is given by

$$Power = T_V\left(\left(\frac{\ln(1-\varepsilon)}{\sigma_W\sqrt{b/n}}\right)-t_{V,1-\alpha}\right)$$

where *T* represents the cumulative *t* distribution, *V* and *b* depend on the design, *n* is the average number of subjects per sequence, and

$$\sigma_W = \sqrt{\ln\left(COV_Y^2+1\right)}$$

The power of the non-inferiority and superiority tests for the case in which higher values are worse is given by

$$Power = 1-T_V\left(t_{V,1-\alpha}-\left(\frac{-\ln(1+\varepsilon)}{\sigma_W\sqrt{b/n}}\right)\right)$$

The constants *V* and *b* depend on the design as follows:

Design Type	**Parameters (*V*,*b*)**
Balaam's Design	*V* = 4n – 3, *b* = 2.
Two-Sequence Dual Design	*V* = 4n – 4, *b* = 3/4.
Four-Period Design with Two Sequences	*V* = 6n – 5, *b* = 11/20.
Four-Period Design with Four Sequences	*V* = 12n – 5, *b* = 1/4.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level. Note that there are two choices for finding *N*. Select *N (Equal Per Sequence)* when you want the design to have an equal number of subjects per sequence. Select *N (Exact)* when you want to find the exact sample size even though the number of subjects cannot be dividing equally among the sequences.

Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in all sequences). These values must be integers greater than one.

Effect Size – Ratios

E (Equivalence Margin)

This is the magnitude of the relative *margin of equivalence*. It is the smallest change in the ratio of the two means that still results in the conclusion of non-inferiority (or superiority).

For example, suppose the non-inferiority boundary for the mean ratio is to be 0.80. This value is interpreted as follows: if the mean ratio (Treatment Mean / Reference Mean) is greater than 0.80, the treatment group is non-inferior to the reference group. In this example, the margin of equivalence would be 1.00 - 0.80 = 0.20.

This example assumed that higher values are better. If higher values are worse, an equivalence margin of 0.20 would be translated into a non-inferiority bound of 1.20. In this case, if the mean ratio is less than 1.20, the treatment group is non-inferior to the reference group.

Note that the sign of this value is ignored. Only the magnitude is used.

Recommended values:

0.20 is a common value for the parameter.

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Usually, the ratio will be assumed to be one. However, some authors recommend calculating the power using a ratio of 1.05 since this will require a larger, more conservative, sample size.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not logged) scale. This value must be determined from past experience or from a pilot study. It is most easily calculated from the within mean-square error of the analysis of variance of the logged data using the relationship

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\,.$$

Test

Design Type

Specify the type of cross-over design that you are analyzing. Note that all of these designs assume that you are primarily interested in the overall difference between the two treatment means.

Test Type

This option specifies the type of test. Select *Non-Inferiority* when you want to test whether the treatment mean is within the margin of equivalence of being no worse than the reference mean.

Select *Superiority* when you want to test whether the treatment mean is better than the reference mean by at least the margin of equivalence.

Higher is

This option defines whether higher values of the response variable are to be considered good or bad. For example, if the response variable is blood pressure, higher values are often considered bad. However, if the response variable is income, higher values are usually considered good.

This option is used with Test Type to determine the direction of the hypothesis test.

Example 1 – Finding Power

A company has developed a generic drug for treating rheumatism and wants to show that it is not inferior to standard drug. Balaam's cross-over design will be used.

Researchers have decided to set the margin of equivalence at 0.20. Past experience leads the researchers to set the COV to 0.40. The significance level is 0.05. The power will be computed assuming that the true ratio is one. Sample sizes between 50 and 550 will be included in the analysis. Note that several of these sample size values are not divisible by 4. This is note a problem here because are main goal is to get an overview of power versus sample size. When searching for the sample size, we can request that only designs divisible by 4 be considered.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**50 to 550 by 100**
E (Equivalence Margin)	**0.20**
R1 (True Ratio)	**1**
COV (Coefficient of Variation)	**0.40**
Design Type	**4x2 (Balaam)**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Annotated Output

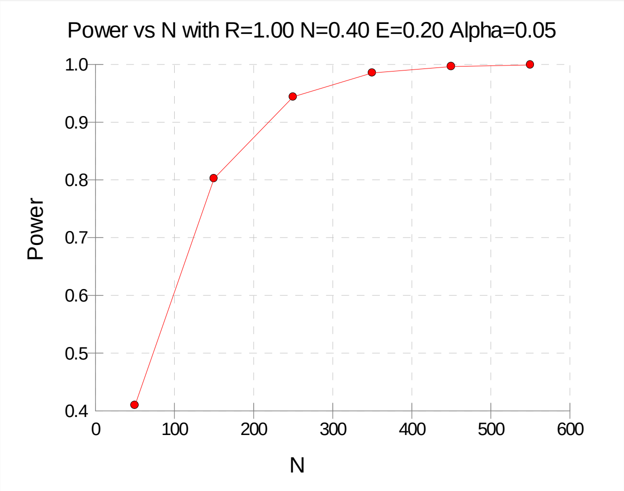
Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Non-Inferiority Using the Mean Ratio
Design: Balaam's Cross-Over.  Hypotheses: H0: R <= 1-E; H1: R > 1-E.

           Total   Sequences                     Mean      Coef.
          Sample         and   Equivalence  Ratio for         of
            Size     Periods        Margin      Power  Variation
Power        (N)       (SxP)           (E)        (R)      (COV)     Alpha      Beta
0.4096        50         4x2          0.20       1.00       0.40    0.0500    0.5904
0.8024       150         4x2          0.20       1.00       0.40    0.0500    0.1976
0.9438       250         4x2          0.20       1.00       0.40    0.0500    0.0562
0.9853       350         4x2          0.20       1.00       0.40    0.0500    0.0147
0.9964       450         4x2          0.20       1.00       0.40    0.0500    0.0036
0.9992       550         4x2          0.20       1.00       0.40    0.0500    0.0008

Report Definitions
H0 (null hypothesis) is that R <= 1-E, where R = Treatment Mean / Reference Mean.
H1 (alternative hypothesis) is that R > 1-E.
Power is the probability of rejecting H0 (concluding non-inferiority) when H0 is false.
N is the total number of subjects. They are divided evenly among all sequences.
E is the magnitude of the relative margin of equivalence.
R is the ratio of the means at which the power is computed.
COV is the coefficient of variation on the original scale.
Alpha is the probability of falsely rejecting H0 (falsely concluding non-inferiority).
Beta is the probability of not rejecting H0 when it is false.
Balaam's Cross-Over Design with pattern: AA; BB; AB; BA

Summary Statements
In a non-inferiority test on data for which higher values are better drawn from Balaam's
cross-over design, a total sample size of 50 achieves 41% power at a 5% significance level when
the true ratio of the means is 1.00, the coefficient of variation is 0.40, and the relative
equivalence margin is 0.20.
```

This report shows the power for the indicated scenarios.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Finding Sample Size

Continuing with Example1, the researchers want to find the exact sample size needed to achieve both 80% and 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.8 0.9**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
E (Equivalence Margin)	**0.20**
R1 (True Ratio)	**1**
COV (Coefficient of Variation)	**0.40**
Design Type	**4x2 (Balaam)**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Non-Inferiority Using the Mean Ratio
Design: Balaam's Cross-Over.  Hypotheses: H0: R <= 1-E; H1: R > 1-E.

             Total  Sequences                   Mean       Coef.
            Sample        and  Equivalence  Ratio for         of
              Size    Periods       Margin      Power  Variation
Power          (N)      (SxP)          (E)        (R)      (COV)     Alpha      Beta
0.9027         208        4x2         0.20       1.00       0.40    0.0500    0.0973
0.8070         152        4x2         0.20       1.00       0.40    0.0500    0.1930
```

When the equivalence margin is set to 0.20, we note that 208 subjects are needed to achieve 90% power and 152 subjects are needed to achieve at least 80% power.

Example 3 – Validation

We could not find a validation example for this procedure in the statistical literature, so we will have to generate a validated example from within ***PASS***. To do this, we use the Higher-Order, Cross-Over Equivalence using Ratios procedure which was validated. By setting the upper equivalence limit to a large value (we used 11), we obtain results for a non-inferiority test that can be used to validate this procedure.

In the other procedure, suppose the coefficient of variation is 0.40, the equivalence limits are 0.80 and 11.0, the true ratio of the means is 1, the power is 90%, and the significance level is 0.05. These settings are stored as Example4 in that procedure. ***PASS*** calculates a sample size of 208.

We will now setup this example in ***PASS***. The only difference is that now we set E to 0.2 instead of RL to 0.8.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Non-Inferiority & Superiority Tests for Two Means in a Higher-Order Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Non-Inferiority & Superiority Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
E (Equivalence Margin)	**0.2**
R1 (True Ratio)	**1.0**
COV (Coefficient of Variation)	**0.40**
Design Type	**4x2 (Balaam)**
Test Type	**Non-Inferiority**
Higher Is	**Good**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing Non-Inferiority Using the Mean Ratio
Design: Balaam's Cross-Over.  Hypotheses: H0: R <= 1-E; H1: R > 1-E.

             Total   Sequences                    Mean      Coef.
            Sample         and  Equivalence  Ratio for         of
              Size     Periods       Margin      Power  Variation
Power          (N)       (SxP)          (E)        (R)      (COV)     Alpha      Beta
0.9027         208         4x2         0.20       1.00       0.40    0.0500    0.0973
```

PASS has also obtained the sample size of 208.

Chapter 540

Equivalence Tests for Two Means in a Higher-Order Cross-Over Design using Differences

Introduction

This procedure calculates power and sample size of statistical tests of equivalence of two means of higher-order cross-over designs when the analysis uses a t-test or equivalent. The parameter of interest is the ratio of the two means. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. Only a brief introduction to the subject will be given here. For a comprehensive discussion on the subject, refer to Chen, Chow, and Li (1997).

Measurements are made on individuals that have been randomly assigned to one of several treatment sequences. This *cross-over* design may be analyzed by a TOST equivalence test to show that the two means do not differ by more than a small amount, called the margin of equivalence.

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments at least once and the object is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to another. It is assumed that there is a *washout* period

between treatments during which the response returns to its baseline value. If this does not occur, there is said to be a *carryover* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence of treatments is treatment A followed by treatment B. The other sequence is B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carryover to the second. Thus, the groups of subjects in this design are defined by the sequence in which the two treatments are administered, not by the treatments they receive.

Higher-Order Cross-Over Designs

Chen et al. (1997) present the results for four cross-over designs that are more complicated than the 2x2 design. Assume that the two treatments are labeled A and B. The available designs are defined by the order and number of times the two treatments are administered.

Balaam’s Design

Balaam’s design has four sequences with two treatments each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2
1	A	A
2	B	B
3	A	B
4	B	A

Two-Sequence Dual Design

This design has two sequences with three periods each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2	Period 3
1	A	B	B
2	B	A	A

Four-Period Design with Two Sequences

This design has two sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	B	B	A
2	B	A	A	B

Four-Period Design with Four Sequences

This design has four sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	A	B	B
2	B	B	A	A
1	A	B	B	A
2	B	A	A	B

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment may be easier to obtain because each patient will receive both treatments.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. In a cross-over experiment, it may be difficult to separate the treatment effect from the time effect and the carry-over effect of the previous treatment.

These cross-over designs cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it cannot be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it may be more difficult to keep patients enrolled in the study. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

Outline of an Equivalence Test

PASS follows the *two one-sided tests* approach described by Schuirmann (1987) and Phillips (1990). It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ε	|E|	*Margin of equivalence*. This is a tolerance value that defines the maximum difference that is not of practical importance. This may be thought of as the largest change from the baseline that is considered to be trivial. The absolute value is shown to emphasize that this is a magnitude. The sign of the value will be determined by the specific design that is being used.
δ	D	*True difference*. This is the value of $\mu_T - \mu_R$, the difference between the treatment and reference means. This is the value at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their difference is needed for power and sample size calculations.

The null hypothesis of non-equivalence is

$$\mathrm{H}_0: \delta \leq \varepsilon_L \text{ or } \delta \geq \varepsilon_U \text{ where } \varepsilon_L < 0, \varepsilon_U > 0.$$

The alternative hypothesis of equivalence is

$$\mathrm{H}_1{:}\, \varepsilon_L < \delta < \varepsilon_U$$

Test Statistics

The analysis for assessing equivalence using higher-order cross-over designs is discussed in detail in Chapter 9 of Chow and Liu (2000). Unfortunately, their presentation is too lengthy to give here. Their method involves the computation of an analysis of variance to estimate the error variance. It also describes the construction of confidence limits for appropriate contrasts. These confidence limits can then be compared to the equivalence limits to test for equivalence. We refer you to their book for details.

Power Calculation

The power is given by

$$Power = T_V\left(\left(\frac{\varepsilon_U - \delta}{\sigma_W \sqrt{b/n}}\right) - t_{V,1-\alpha}\right) - T_V\left(t_{V,1-\alpha} - \left(\frac{\delta - \varepsilon_L}{\sigma_W \sqrt{b/n}}\right)\right)$$

where *T* represents the cumulative *t* distribution, *V* and *b* depend on the design, σ_W is the square root of the within mean square error from the ANOVA table used to analyze the cross-over design, and *n* is the average number of subjects per sequence. Note that the constants *V* and *b* depend on the design as follows.

Design Type	**Parameters (*V*,*b*)**
Balaam's Design	*V* = 4n – 3, *b* = 2.
Two-Sequence Dual Design	*V* = 4n – 4, *b* = 3/4.
Four-Period Design with Two Sequences	*V* = 6n – 5, *b* = 11/20.
Four-Period Design with Four Sequences	*V* = 12n – 5, *b* = 1/4.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level. Note that there are two choices for finding *N*. Select *N (Equal Per Sequence)* when you want the design to have an equal number of subjects per sequence. Select *N (Exact)* when you want to find the exact sample size even though the number of subjects cannot be dividing equally among the sequences.

Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in all sequences). This value must be an integer greater than one.

You may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Equivalence Limits

|EU| (Upper Equivalence Limit)

This value gives the upper limit of equivalence. Differences outside EL and EU are not considered equivalent, while differences between them are.

Note that EL must be less than zero and EU must be greater than zero. Also, D, EL, and EU must satisfy EL<D<EU. Finally, the scale of these numbers must match the scale of Sw.

-|EL| (Lower Equivalence Limit)

This value gives lower limit on equivalence. Differences outside EL and EU are not considered equivalent, while differences between them are.

If you want symmetric limits, enter -UPPER LIMIT for EL to force EL = -|EU|.

Note that EL must be less than zero and EU must be greater than zero. Also, D, EL, and EU must satisfy EL<D<EU. Finally, the scale of these numbers must match the scale of Sw.

Effect Size – True Mean Difference

D (True Difference)

This is the true difference between the two means at which the power is to be computed. Often this value is set to zero, but it can be non-zero as long as it is between the equivalence limits, EL and EU.

D, EL, and EU must satisfy EL<D<EU. Finally, the scale of these numbers must match the scale of Sw.

Effect Size – Standard Deviation

Sw (Within Standard Error)

Specify one or more values of Sw, which is SQR(WMSE) where WMSE is the within mean square error from the ANOVA table used to analyze the cross-over design. These values must be positive.

You can press the SD button to load the Standard Deviation Estimator window.

Test

Design Type

Specify the type of cross-over design that you are analyzing. Note that all of these designs assume that you are primarily interested in the overall difference between the two treatment means.

Example 1 – Finding Power

A two-sequence, dual cross-over design is to be used to compare the impact of two drugs on diastolic blood pressure. The average diastolic blood pressure after administration of the reference drug is 96 mmHg. Researchers believe this average may drop to 92 mmHg with the use of a new drug. The within mean square error found from similar studies is 324. Its square root is 18.

Following FDA guidelines, the researchers want to show that the diastolic blood pressure is within 20% of the diastolic blood pressure of the reference drug. Thus, the equivalence limits of the mean difference of the two drugs are -19.2 and 19.2. They decide to calculate the power for a range of sample sizes between 4 and 40. The significance level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a Higher-Order Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)............................	**4 6 8 10 12 14 16 18 20 30 40**
|EU| (Upper Equivalence Limit)..............	**19.2**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**-4**
Sw (Within Standard Error)	**18**
Design Type	**2x3 (Two-Sequence Dual)**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

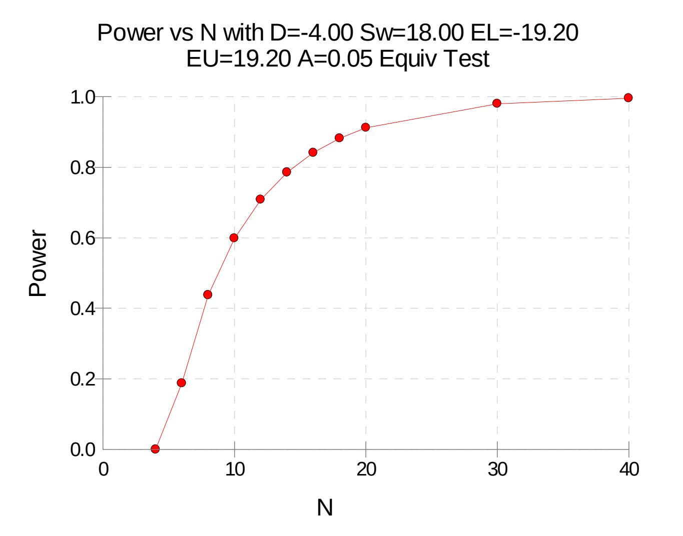
```
Numeric Results for Testing the Equivalence of Two Means
Design: Two-Sequence Dual Cross-Over

              Total  Sequences   Lower    Upper   Diff.  Standard
             Sample        and  Equiv.   Equiv.     for     Error
               Size    Periods   Limit    Limit   Power   of Diff.
Power           (N)      (SxP)    (EL)     (EU)     (D)      (Sw)    Alpha     Beta
0.0000            4        2x3  -19.20    19.20   -4.00     18.00   0.0500   1.0000
0.1878            6        2x3  -19.20    19.20   -4.00     18.00   0.0500   0.8122
0.4375            8        2x3  -19.20    19.20   -4.00     18.00   0.0500   0.5625
0.5985           10        2x3  -19.20    19.20   -4.00     18.00   0.0500   0.4015
Report Continues…
```

Report Definitions
Power is the probability of rejecting non-equivalence when the means are equivalent.
N is the total number of subjects. They are divided evenly among all sequences.
S is the number of sequences.
P is the number of periods per sequence.
EU & EL are the upper & lower limits of the maximum allowable difference that results in equivalence.
D is the difference between the means at which the power is computed.
Sw is the square root of the within mean square error from the ANOVA table.
Alpha is the probability of rejecting non-equivalence when they are non-equivalent.
Beta is the probability of accepting non-equivalence when they are equivalent.
Two-Sequence Dual Cross-Over Design with pattern: ABB; BAA

Summary Statements
In an equivalence test of means using two one-sided tests on data from a two-sequence dual cross-over design, a total sample size of 4 achieves 0% power at a 5% significance level when the true difference between the means is -4.00, the square root of the within mean square error is 18.00, and the equivalence limits are -19.20 and 19.20.

This report shows the power for the indicated scenarios.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Finding Sample Size

Continuing with Example1, the researchers want to find the exact sample size needed to achieve both 80% and 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a Higher-Order Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.80 0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
|EU| (Upper Equivalence Limit)	**19.2**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**-4**
Sw (Within Standard Error)	**18**
Design Type	**2x3 (Two-Sequence Dual)**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing the Equivalence of Two Means
Design: Two-Sequence Dual Cross-Over

           Total  Sequences    Lower     Upper    Diff.  Standard
          Sample        and   Equiv.    Equiv.      for     Error
            Size    Periods    Limit     Limit    Power  of Diff.
Power        (N)      (SxP)     (EL)      (EU)      (D)      (Sw)     Alpha      Beta
0.9119        20        2x3   -19.20     19.20    -4.00     18.00    0.0500    0.0881
0.8411        16        2x3   -19.20     19.20    -4.00     18.00    0.0500    0.1589
```

Twenty subjects are needed to achieve at least 90% power and sixteen subjects are needed to achieve at least 80% power.

Example 3 – Validation using Chen

Chen et al. (1997) page 757 present a table of sample sizes for various parameter values. In this table, the treatment mean, standard deviation, and equivalence limits are all specified as percentages of the reference mean. We will reproduce the seventeenth line of the table in which the square root of the within mean square error is 10%, the equivalence limits are 20%, the difference between the means is 0%, 5%, 10%, and 15%, the power is 90%, and the significance level is 0.05. Chen reports total sample sizes of 24, 36, 72, and 276. We will now setup this example in ***PASS***.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a Higher-Order Cross-Over Design [Differences]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Differences**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
|EU| (Upper Equivalence Limit)	**0.2**
-|EL| (Lower Equivalence Limit)	**-Upper Limit**
D (True Difference)	**0 0.05 0.10 0.15**
Sw (Within Standard Error)	**0.1**
Design Type	**4x2 (Balaam)**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing the Equivalence of Two Means
Design: Balaam's Cross-Over

           Total  Sequences   Lower    Upper    Diff.  Standard
          Sample        and  Equiv.   Equiv.     for     Error
            Size    Periods   Limit    Limit   Power   of Diff.
Power        (N)      (SxP)    (EL)     (EU)     (D)      (Sw)    Alpha     Beta
0.9041        24        4x2   -0.20     0.20    0.00      0.10   0.0500   0.0959
0.9266        36        4x2   -0.20     0.20    0.05      0.10   0.0500   0.0734
0.9065        72        4x2   -0.20     0.20    0.10      0.10   0.0500   0.0935
0.9003       276        4x2   -0.20     0.20    0.15      0.10   0.0500   0.0997
```

PASS obtains the same samples sizes as Chen et al. (1997).

Chapter 545

Equivalence Tests for Two Means in a Higher-Order Cross-Over Design using Ratios

Introduction

This procedure calculates power and sample size of statistical tests of equivalence of two means of higher-order cross-over designs when the analysis uses a t-test or equivalent. The parameter of interest is the ratio of the two means. Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. Only a brief introduction to the subject will be given here. For a comprehensive discussion on the subject, refer to Chen, Chow, and Li (1997).

Measurements are made on individuals that have been randomly assigned to one of several treatment sequences. This *cross-over* design may be analyzed by a TOST equivalence test to show that the two means do not differ by more than a small amount, called the margin of equivalence.

Measurements are made on individuals that have been randomly assigned to one of several treatment sequences. This *cross-over* design may be analyzed by a TOST equivalence test to show that the two means do not differ by more than a small amount, called the margin of equivalence.

Cross-Over Designs

Senn (2002) defines a *cross-over* design as one in which each subject receives all treatments at least once and the object is to study differences among the treatments. The name *cross-over* comes from the most common case in which there are only two treatments. In this case, each subject *crosses over* from one treatment to another. It is assumed that there is a *washout* period between treatments during which the response returns to its baseline value. If this does not occur, there is said to be a *carryover* effect.

A 2x2 cross-over design refers to two treatments (periods) and two *sequences* (treatment orderings). One sequence of treatments is treatment A followed by treatment B. The other sequence is B and then A. The design includes a washout period between responses to make certain that the effects of the first drug do no carryover to the second. Thus, the groups of subjects in this design are defined by the sequence in which the two treatments are administered, not by the treatments they receive.

Higher-Order Cross-Over Designs

Chen et al. (1997) present the results for four cross-over designs that are more complicated than the 2x2 design. Assume that the two treatments are labeled A and B. The available designs are defined by the order and number of times the two treatments are administered.

Balaam's Design

Balaam's design has four sequences with two treatments each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2
1	A	A
2	B	B
3	A	B
4	B	A

Two-Sequence Dual Design

This design has two sequences with three periods each. It is popular because it allows the intrasubject variabilities to be estimated. The design is

Sequence	Period 1	Period 2	Period 3
1	A	B	B
2	B	A	A

Four-Period Design with Two Sequences

This design has two sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	B	B	A
2	B	A	A	B

Four-Period Design with Four Sequences

This design has four sequences of four periods each. The design is

Sequence	Period 1	Period 2	Period 3	Period 4
1	A	A	B	B
2	B	B	A	A
1	A	B	B	A
2	B	A	A	B

Advantages of Cross-Over Designs

A comparison of treatments on the same subject is expected to be more precise. The increased precision often translates into a smaller sample size. Also, patient enrollment may be easier to obtain because each patient will receive both treatments.

Disadvantages of Cross-Over Designs

The statistical analysis of a cross-over experiment is more complex than a parallel-group experiment and requires additional assumptions. In a cross-over experiment, it may be difficult to separate the treatment effect from the time effect and the carry-over effect of the previous treatment.

These cross-over designs cannot be used when the treatment (or the measurement of the response) alters the subject permanently. Hence, it cannot be used to compare treatments that are intended to provide a cure.

Because subjects must be measured at least twice, it may be more difficult to keep patients enrolled in the study. This is particularly true when the measurement process is painful, uncomfortable, embarrassing, or time consuming.

Outline of an Equivalence Test

PASS follows the *two one-sided tests* approach described by Schuirmann (1987) and Phillips (1990). It will be convenient to adopt the following specialized notation for the discussion of these tests.

Parameter	PASS Input/Output	Interpretation
μ_T	Not used	*Treatment mean*. This is the treatment mean.
μ_R	Not used	*Reference mean*. This is the mean of a reference population.
ϕ_L, ϕ_U	RL, RU	*Margin of equivalence*. These limits define an interval of the ratio of the means in which their difference is so small that it may be ignored.
ϕ	R1	*True ratio*. This is the value of $\phi = \mu_T / \mu_R$ at which the power is calculated.

Note that the actual values of μ_T and μ_R are not needed. Only their ratio is needed for power and sample size calculations.

The null hypothesis of non-equivalence is

$$\mathrm{H}_0{:}\, \phi \le \phi_L \text{ or } \phi \ge \phi_U \text{ where } \phi_L < 1,\ \phi_U > 1.$$

The alternative hypothesis of equivalence is

$$\mathrm{H}_1{:}\, \phi_L < \phi < \phi_U$$

Log-Transformation

In many cases, hypotheses stated in terms of ratios are more convenient than hypotheses stated in terms of differences. This is because ratios can be interpreted as scale-less percentages, but differences must be interpreted as actual amounts in their original scale. Hence, it has become a common practice to take the following steps in hypothesis testing.

1. State the statistical hypothesis in terms of a ratio.
2. Transform this into a hypothesis about the difference by taking logarithms.
3. Analyze the logged data—that is, do the analysis in terms of the difference.
4. Draw the conclusion in terms of the ratio.

The details of step 2 for the alternative hypothesis are as follows.

$$\begin{aligned}
&\phi_L < \phi < \phi_U \\
&\Rightarrow \phi_L < \left\{ \frac{\mu_T}{\mu_R} \right\} < \phi_U \\
&\Rightarrow \ln(\phi_L) < \left\{ \ln(\mu_T) - \ln(\mu_R) \right\} < \ln(\phi_U)
\end{aligned}$$

Thus, a hypothesis about the ratio of the means on the original scale can be translated into a hypothesis about the difference of two means on the logged scale.

When performing an equivalence test on the difference between means, the usual procedure is to set the equivalence limits symmetrically above and below zero. Thus, the equivalence limits will be plus or minus an appropriate amount. The common practice is to do the same when the data are being analyzed on the log scale. However, when symmetric limits are set on the log scale, they do not translate to symmetric limits on the original scale. Instead, they translate to limits that are the inverses of each other.

Perhaps these concepts can best be understood by considering an example. Suppose the researchers have determined that the lower equivalence limit should be 80% on the original scale. Since they are planning to use a log scale for their analysis, they transform this limit into the log scale by taking the logarithm of 0.80. The result is -0.223144. Wanting symmetric limits, they set the upper equivalence limit to 0.223144. Exponentiating this value, they find that exp(0.223144) = 1.25. Note that 1/(0.80) = 1.25. Thus, the limits on the original scale are 80% and 125%, not 80% and 120%.

Using this procedure, appropriate equivalence limits for the ratio of two means can be easily determined. Here are a few sets of equivalence limits for ratios.

	Lower	Upper	Lower	Upper
Specified	Limit	Limit	Limit	Limit
Percent	Original	Original	Log	Log
Change	Scale	Scale	Scale	Scale
-25%	75.0%	133.3%	-0.287682	0.287682
+25%	80.0%	125.0%	-0.223144	0.223144
-20%	80.0%	125.0%	-0.223144	0.223144
+20%	83.3%	120.0%	-0.182322	0.182322
-10%	90.0%	111.1%	-0.105361	0.105361
+10%	90.9%	110.0%	-0.095310	0.095310

Note that negative percent-change values specify the lower limit first, while positive percent-change values specify the upper limit first. After the first limit is found, the other limit is calculated as its inverse.

Coefficient of Variation

The coefficient of variation (COV) is the ratio of the standard deviation to the mean. This parameter can be used to represent the variation in the data because of a unique relationship that it has in the case of log-normal data.

Suppose the variable *X* is the logarithm of the original variable *Y*. That is, *X* = ln(*Y*) and *Y* = exp(*X*). Label the mean and variance of *X* as μ_X and σ_X^2, respectively. Similarly, label the mean and variance of *Y* as μ_Y and σ_Y^2, respectively. If *X* is normally distributed, then *Y* is log-normally distributed. Julious (2004) presents the following well-known relationships between these two variables

$$\mu_Y = \left(e^{\mu_X + \frac{\sigma_X^2}{2}} \right)$$

$$\sigma_Y^2 = \mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)$$

From this relationship, the coefficient of variation of *Y* can be expressed as

$$\begin{aligned} COV_Y &= \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y} \\ &= \sqrt{e^{\sigma_X^2} - 1} \end{aligned}$$

Solving this relationship for σ_X^2, the standard deviation of *X* can be stated in terms of the coefficient of variation of *Y*. This equation is

$$\sigma_X = \sqrt{\ln\left(COV_Y^2 + 1 \right)}$$

Similarly, the mean of *X* is

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

Thus, the hypotheses can be stated in the original (Y) scale and then the power can be analyzed in the transformed (X) scale.

Test Statistics

The analysis for assessing equivalence using higher-order cross-over designs is discussed in detail in Chapter 9 of Chow and Liu (2000). Unfortunately, their presentation is too lengthy to give here. Their method involves the computation of an analysis of variance to estimate the error variance. It also describes the construction of confidence limits for appropriate contrasts. These confidence limits can then be compared to the equivalence limits to test for equivalence. We refer you to their book for details.

Power Calculation

The power is given by

$$Power = T_V\left(\left(\frac{\ln(\phi_U) - \left|\ln(\phi)\right|}{\sigma_W\sqrt{b/n}}\right) - t_{V,1-\alpha}\right) - T_V\left(t_{V,1-\alpha} - \left(\frac{-\ln(\phi_L) + \left|\ln(\phi)\right|}{\sigma_W\sqrt{b/n}}\right)\right)$$

where

$$\sigma_W = \sqrt{\ln\left(COV_Y^2 + 1\right)},$$

T represents the cumulative *t* distribution, *V* and *b* depend on the design, and *n* is the average number of subjects per sequence. Note that the constants *V* and *b* depend on the design as follows.

Design Type	**Parameters (*V*,*b*)**
Balaam’s Design	*V* = 4n – 3, *b* = 2.
Two-Sequence Dual Design	*V* = 4n – 4, *b* = 3/4.
Four-Period Design with Two Sequences	*V* = 6n – 5, *b* = 11/20.
Four-Period Design with Four Sequences	*V* = 12n – 5, *b* = 1/4.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. Under most situations, you will select either *Power and Beta* for a power analysis or *N* for sample size determination.

Select *N* when you want to calculate the sample size needed to achieve a given power and alpha level. Note that there are two choices for finding *N*. Select *N (Equal Per Sequence)* when you want the design to have an equal number of subjects per sequence. Select *N (Exact)* when you want to find the exact sample size even though the number of subjects cannot be dividing equally among the sequences.

Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Total Sample Size)

This option specifies one or more values of the sample size, the number of individuals in the study (total subjects in all sequences). This value must be an integer greater than one.

You may enter a list of values using the syntax *50,100,150,200,250* or *50 to 250 by 50*.

Effect Size – Equivalence Limits

RU (Upper Equivalence Limit)

Enter the upper equivalence limit for the ratio of the two means. When the ratio of the means is between this value and RL, the two means are said to be equivalent. The value must be greater than one. A popular choice is 1.25. Note that this value is not a percentage.

If you enter *1/RL*, then 1/RL will be calculated and used here. This choice is commonly used because RL and 1/RL give limits that are of equal magnitude on the log scale.

RL (Lower Equivalence Limit)

Enter the lower equivalence limit for the ratio of the two means. When the ratio of the means is between this value and RU, the two means are said to be equivalent. The value must be less than one. A popular choice is 0.80. Note that this value is not a percentage.

If you enter *1/RU*, then 1/RU will be calculated and used here. This choice is commonly used because RU and 1/RU give limits that are of equal magnitude on the log scale.

Effect Size – True Ratio

R1 (True Ratio)

This is the value of the ratio of the two means at which the power is to be calculated. Usually, the ratio will be assumed to be one. However, some authors recommend calculating the power using a ratio of 1.05 since this will require a larger, more conservative, sample size.

Effect Size – Coefficient of Variation

COV (Coefficient of Variation)

The coefficient of variation is used to specify the variability (standard deviation). It is important to realize that this is the COV defined on the original (not logged) scale. This value must be determined from past experience or from a pilot study. It is most easily calculated from the within mean-square error of the analysis of variance of the logged data using the relationship

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1} .$$

Test

Design Type

Specify the type of cross-over design that you are analyzing. Note that all of these designs assume that you are primarily interested in the overall difference between two means.

Example 1 – Finding Power

A company has opened a new manufacturing plant and wants to show that the drug produced in the new plant is equivalent to that produced in the older plant. A two-sequence, dual cross-over design will be used to test the equivalence of drugs produced at the two plants.

Researchers have decided to set the equivalence limits for the ratio at 0.80 and 1.25. Past experience leads the researchers to set the COV to 0.40. The significance level is 0.05. The power will be computed assuming that the true ratio is 0.96. Sample sizes between 10 and 80 will be included in the analysis.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a Higher-Order Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N (Total Sample Size)	**10 20 30 40 60 80**
RU (Upper Equivalence Limit)	**1.25**
RL (Lower Equivalence Limit)	**1/RU**
R1 (True Ratio)	**0.96**
COV (Coefficient of Variation)	**0.40**
Design Type	**2x3 (Two-Sequence Dual)**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

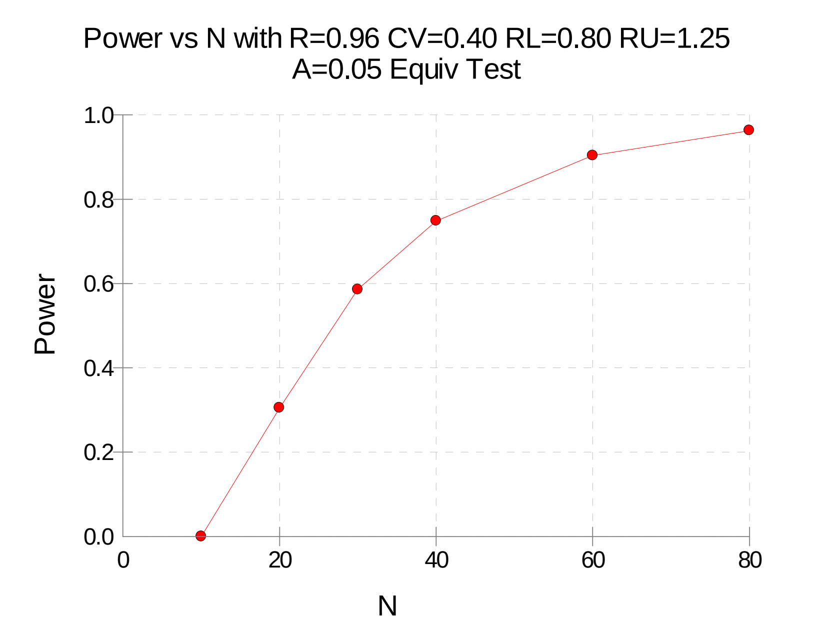
```
           Total   Sequences    Lower     Upper     Mean      Coef.
          Sample         and   Equiv.    Equiv. Ratio for        of
            Size     Periods    Limit     Limit    Power  Variation
Power        (N)       (SxP)     (RL)      (RU)     (R1)      (COV)     Alpha     Beta
0.0000        10         2x3     0.80      1.25     0.96       0.40    0.0500   1.0000
0.3051        20         2x3     0.80      1.25     0.96       0.40    0.0500   0.6949
0.5858        30         2x3     0.80      1.25     0.96       0.40    0.0500   0.4142
0.7483        40         2x3     0.80      1.25     0.96       0.40    0.0500   0.2517
0.9035        60         2x3     0.80      1.25     0.96       0.40    0.0500   0.0965
0.9627        80         2x3     0.80      1.25     0.96       0.40    0.0500   0.0373
```

Report Definitions
Power is the probability of rejecting non-equivalence when the means are equivalent.
N is the total number of subjects. They are divided evenly among all sequences.
RU & RL are the upper and lower equivalence limits. Ratios between these limits are equivalent.
R1 is the ratio of the means at which the power is computed.
COV is the coefficient of variation on the original scale.
Alpha is the probability of rejecting non-equivalence when they are non-equivalent.
Beta is the probability of accepting non-equivalence when they are equivalent.
Two-Sequence Dual Cross-Over Design with pattern: ABB; BAA

Summary Statements
In an equivalence test of means using two one-sided tests on data from a two-sequence dual cross-over design, a total sample size of 10 achieves 0% power at a 5% significance level when the true ratio of the means is 0.96, the coefficient of variation on the original (unlogged) scale is 0.40, and the equivalence limits are 0.80 and 1.25.

This report shows the power for the indicated scenarios. Note that 60 subjects will yield a power of just over 90%.

Plots Section



This plot shows the power versus the sample size.

Example 2 – Finding Sample Size

Continuing with Example 1, the researchers want to find the exact sample size needed to achieve both 80% power and 90% power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a Higher-Order Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal per Sequence)**
Power	**0.80 0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
RU (Upper Equivalence Limit)	**1.25**
RL (Lower Equivalence Limit)	**1/RU**
R1 (True Ratio)	**0.96**
COV (Coefficient of Variation)	**0.40**
Design Type	**2x3 (Two-Sequence Dual)**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
          Total  Sequences   Lower   Upper     Mean     Coef.
         Sample        and  Equiv.  Equiv. Ratio for       of
           Size    Periods   Limit   Limit    Power Variation
 Power      (N)      (SxP)    (RL)    (RU)     (R1)     (COV)    Alpha     Beta
 0.9035      60        2x3    0.80    1.25     0.96      0.40   0.0500   0.0965
 0.8119      46        2x3    0.80    1.25     0.96      0.40   0.0500   0.1881
```

We note that 60 subjects are needed to achieve 90% power and 46 subjects are needed to achieve at least 80% power.

Example 3 – Validation using Chen

Chen et al. (1997) page 761 presents a table of sample sizes for various parameter values for Balaam’s design. We will reproduce entries from the first and seventeenth lines of the table in which the COV is 10%, the equivalence limits are 0.8 and 1.25, the actual ratio of between the means is 1, the power values are 80% and 90%, and the significance level is 0.05. Chen reports total sample sizes of 16 and 20. We will now setup this example in ***PASS***.

The COV entered by Chen is the COV of the logged data. Since ***PASS*** requires the COV of the original data, we must use the relationship

$$\begin{aligned} COV_Y &= \sqrt{e^{\sigma_w^2} - 1} \\ &= \sqrt{e^{0.1^2} - 1} \\ &= \sqrt{e^{0.01} - 1} \\ &= 0.10025 \end{aligned}$$

to obtain the appropriate value of COV.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Equivalence Tests for Two Means in a Higher-Order Cross-Over Design [Ratios]** procedure window by clicking on **Means**, then **Two Means**, then **Higher-Order Cross-Over Designs**, then **Equivalence Tests**, then **Specify using Ratios**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Equal Per Sequence)**
Power	**0.80 0.90**
Alpha	**0.05**
N (Total Sample Size)	*Ignored since this is the Find setting*
RU (Upper Equivalence Limit)	**1.25**
RL (Lower Equivalence Limit)	**1/RU**
R1 (True Ratio)	**1**
COV (Coefficient of Variation)	**0.10025**
Design Type	**4x2 (Balaam)**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing the Equivalence of Two Means Using Ratios
Design: Balaam's Cross-Over

            Total  Sequences     Lower     Upper     Mean       Coef.
           Sample        and    Equiv.    Equiv.  Ratio for        of
             Size    Periods     Limit     Limit     Power  Variation
Power         (N)      (SxP)      (RL)      (RU)       (R)      (COV)     Alpha      Beta
0.9085         20        4x2      0.80      1.25      1.00       0.10    0.0500    0.0915
0.8106         16        4x2      0.80      1.25      1.00       0.10    0.0500    0.1894
```

Note that ***PASS*** has obtained the same samples sizes as Chen et al. (1997).

Chapter 550

One-Way Analysis of Variance

Introduction

A common task in research is to compare the averages of two or more populations (groups). We might want to compare the income level of two regions, the nitrogen content of three lakes, or the effectiveness of four drugs. The one-way analysis of variance compares the means of two or more groups to determine if at least one mean is different from the others. The *F* test is used to determine statistical significance. *F* tests are non-directional in that the null hypothesis specifies that all means are equal and the alternative hypothesis simply states that at least one mean is different.

The methods described here are usually applied to the one-way experimental design. This design is an extension of the design used for the two-sample *t* test. Instead of two groups, there are three or more groups. With careful modifications, this procedure may be used to test interaction terms as well.

Planned Comparisons

PASS performs power and sample size calculations for user-specified contrasts.

The usual *F* test tests the hypothesis that all means are equal versus the alternative that at least one mean is different from the rest. Often, a more specific alternative is desired. For example, you might want to test whether the treatment means are different from the control mean, the low dose is different from the high dose, a linear trend exists across dose levels, and so on. These questions are tested using planned comparisons.

We call the comparison *planned* because it was determined before the experiment was conducted. We planned to test the comparison.

A comparison is a weighted average of the means, in which the weights may be negative. When the weights sum to zero, the comparison is called a *contrast*. ***PASS*** provides results for contrasts. To specify a contrast, we need only specify the weights. Statisticians call these weights the *contrast coefficients*.

For example, suppose an experiment conducted to study a drug will have three dose levels: none (control), 20 mg., and 40 mg. The first question is whether the drug made a difference. If it did, the average response for the two groups receiving the drug should be different from the control. If we label the group means M0, M20, and M40, we are interested in comparing M0 with M20 and M40. This can be done in two ways. One way is to construct two tests, one comparing M0 and

M20 and the other comparing M0 and M40. Another method is to perform one test comparing M0 with the average of M20 and M40. These tests are conducted using planned comparisons. The coefficients are as follows:

M0 vs. M20

To compare M0 versus M20, use the coefficients -1,1,0. When applied to the group means, these coefficients result in the comparison M0(-1)+M20(1)+M40(0) which reduces to M20-M0. That is, this contrast results in the difference between the two group means. We can test whether this difference is non-zero using the *t* test (or *F* test since the square of the *t* test is an *F* test).

M0 vs. M40

To compare M0 versus M40, use the coefficients -1,0,1. When applied to the group means, these coefficients result in the comparison M0(-1)+M20(0)+M40(1) which reduces to M40-M0. That is, this contrast results in the difference between the two group means.

M0 vs. Average of M20 and M40

To compare M0 versus the average of M20 and M40, use the coefficients -2,1,1. When applied to the group means, these coefficients result in the comparison M0(-2)+M20(1)+M40(1) which is equivalent to M40+M20-2(M0).

To see how these coefficients were obtained, consider the following manipulations. Beginning with the difference between the average of M20 and M40 and M0, we obtain the coefficients above—aside from a scale factor of one-half.

$$\begin{aligned}\frac{M20+M40}{2}-M0 &= \frac{M20}{2}+\frac{M40}{2}-\frac{M0}{1}\\ &= \frac{1}{2}M20+\frac{1}{2}M40-M0\\ &= \frac{1}{2}\left(M20+M40-2M0\right)\end{aligned}$$

Assumptions

Using the *F* test requires certain assumptions. One reason for the popularity of the *F* test is its robustness in the face of assumption violation. However, if an assumption is not even approximately met, the significance levels and the power of the *F* test are invalidated. Unfortunately, in practice it often happens that several assumptions are not met. This makes matters even worse. Hence, steps should be taken to check the assumptions before important decisions are made.

The assumptions of the one-way analysis of variance are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution. Each group is normally distributed about the group mean.
3. The variances of the populations are equal.

4. The groups are independent. There is no relationship among the individuals in one group as compared to another.
5. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

Technical Details for the One-Way ANOVA

Suppose *k* groups each have a normal distribution and equal means $\left(\mu_1 = \mu_2 = \cdots = \mu_k\right)$. Let $n_1 = n_2 = \cdots = n_k$ denote the number of subjects in each group and let *N* denote the total sample size of all groups. Let $\bar{\mu}_W$ denote the weighted mean of all groups. That is

$$\bar{\mu}_W = \sum_{i=1}^{k} \left(\frac{n_i}{N}\right) \mu_i$$

Let σ denote the common standard deviation of all groups.

Given the above terminology, the ratio of the mean square between groups to the mean square within groups follows a central *F* distribution with two parameters matching the degrees of freedom of the numerator mean square and the denominator mean square. When the null hypothesis of mean equality is rejected, the above ratio has a noncentral *F* distribution which also depends on the noncentrality parameter, λ. This parameter is calculated as

$$\lambda = N \frac{\sigma_m^2}{\sigma^2}$$

where

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{n_i \left(\mu_i - \bar{\mu}_W\right)^2}{N}}$$

Some authors use the symbol ϕ for the noncentrality parameter. The relationship between the two noncentrality parameters is

$$\phi = \sqrt{\frac{\lambda}{k}}.$$

The process of planning an experiment should include the following steps:

1. Determine an estimate of the within group standard deviation, σ. This may be done from prior studies, from experimentation with the Standard Deviation Estimation module, from pilot studies, or from crude estimates based on the range of the data. See the chapter on estimating the standard deviation for more details.
2. Determine a set of means that represent the group differences that you want to detect.
3. Determine the appropriate group sample sizes that will ensure desired levels of α and β. Although it is tempting to set all group sample sizes equal, it is easy to show that putting more subjects in some groups than in others may have better power than keeping group sizes equal (see Example 4).

Power Calculations for One-Way ANOVA

The calculation of the power of a particular test proceeds as follows:

1. Determine the critical value, $F_{k-1,N-k,\alpha}$ where α is the probability of a type-I error and k and N are defined above. Note that this is a two-tailed test as no direction is assigned in the alternative hypothesis.
2. From a hypothesized set of μ_i's, calculate the noncentrality parameter λ based on the values of N, k, σ_m, and σ.
3. Compute the power as the probability of being greater than $F_{k-1,N-k,\alpha}$ on a noncentral-F distribution with noncentrality parameter λ.

Technical Details for a Planned Comparison

The terminology of planned comparisons is identical to that of the one-way AOV, so the notation used above will be repeated here.

Suppose you want to test whether the contrast C

$$C = \sum_{i=1}^{k} c_i \mu_i$$

is significantly different from zero. Here the c_i's are the contrast coefficients.

Define

$$\sigma_{mc} = \frac{\left|\sum_{i=1}^{k} c_i \mu_i\right|}{\sqrt{N \sum_{i=1}^{k} \frac{c_i^2}{n_i}}}$$

Define the noncentrality parameter λ_c, as

$$\lambda_C = N \frac{\sigma_{mc}^2}{\sigma^2}$$

Power Calculations for Planned Comparisons

The calculation of the power of a particular test proceeds as follows:

1. Determine the critical value, $F_{1,N-k,\alpha}$ where α is the probability of a type-I error and k and N are defined above. Note that this is a two-tailed test as no direction is assigned in the alternative hypothesis.
2. From a hypothesized set of μ_i's, calculate the noncentrality parameter λ_c based on the values of N, k, σ_{mc}, and σ.

3. Compute the power as the probability of being greater than $F_{1,N-k,\alpha}$ on a noncentral-*F* distribution with noncentrality parameter λ_c .

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *SM*, *S*, *k*, *n*, *Alpha*, and *Power and Beta*. Under most situations, you will select either *Power and Beta* for a power analysis or *n* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size / Groups – Sample Size Multiplier

n (Sample Size Multiplier)

This is the base, per group, sample size. One or more values, separated by blanks or commas, may be entered. A separate analysis is performed for each value listed here.

The group samples sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by *m1*, *m2*, *m3*, …, *mk* and this value is represented by *n*, the group sample sizes *N1*, *N2*, *N3*, ..., *Nk* are calculated as follows:

N1=[n(m1)]

N2=[n(m2)]

N3=[n(m3)]

etc.

where the operator, [*X*] means the next integer after *X*, e.g. [3.1]=4.

For example, suppose there are three groups and the Group Sample Size Pattern is set to *1,2,3*. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to *2,4,6,8,10*, five sets of group sample sizes will be generated and an analysis run for each. These sets are:

2	4	6
4	8	12
6	12	18
8	16	24
10	20	30

As a second example, suppose there are three groups and the Group Sample Size Pattern is *0.2,0.3,0.5*. When the fractional Pattern values sum to one, n can be interpreted as the total sample size of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

(0.2)12 = 2.4 which is rounded up to the next whole integer, 3.

(0.3)12 = 3.6 which is rounded up to the next whole integer, 4.

(0.5)12 = 6.

Note that in this case, 3+4+6 does not equal n (which is 12). This can happen because of rounding.

Sample Size / Groups – Groups

k (Number of Groups)

This is the number of group means being compared. It must be greater than or equal to two.

You can enter a list of values, in which case, a separate analysis will be calculated for each value. Commas or blanks may separate the numbers. A TO-BY list may be used.

Note that the number of items used in the Hypothesized Means box and the Group Sample Size Pattern box is controlled by this number.

Examples:

2,3,4

2 3 4

2 to 10 by 2

Group Sample Size Pattern

A set of positive, numeric values, one for each group, is entered here. The sample size of group *i* is found by multiplying the ith number from this list times the value of n and rounding up to the next whole number. The number of values must match the number of groups, *k*. When too few numbers are entered, 1's are added. When too many numbers are entered, the extras are ignored.

- **Equal**

  If all sample sizes are to be equal, enter "Equal" here and the desired sample size in n. A set of *k* 1's will be used. This will result in *N1* = *N2* = *N3* = *n*. That is, all sample sizes are equal to *n*.

Effect Size – Means

Hypothesized Means

Enter a set of hypothesized means, one for each group. These means represent the group centers under the alternative hypothesis (the null hypothesis is that they are equal). The standard deviation of these means (*SM*) is used in the power calculations to represent the average size of the differences among the means. The standard deviation of the means is calculated using the formula:

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{(\mu_i - \bar{\mu})^2}{k}}$$

This quantity gives the magnitude of the differences among the group means. Note that when all means are equal, σ_m is zero.

You should enter a set of means that give the pattern of differences you expect or the pattern that you wish to detect. For example, in a particular study involving three groups, your research might be "meaningful" if either of two treatment means is 50% larger than the control mean. If the control mean is 50, then you would enter *50,75,75* as the three means.

It is usually more intuitive to enter a set of mean values. However, it is possible to enter the standard deviation of the means directly by placing an *S* in front of the number (see below).

Some might wish to specify the alternative hypothesis as the effect size, *f*, which is defined as

$$f = \frac{\sigma_m}{\sigma}$$

If so, set $\sigma = 1$ and $\sigma_m = f$. Cohen (1988) has designated values of *f* less than 0.1 as *small*, values around 0.25 to be *medium*, and values over 0.4 to be *large*.

Entering a List of Means

If a set of numbers is entered without a leading *S*, they are assumed to be the hypothesized group means under the alternative hypothesis. Their standard deviation will be calculated and used in the calculations. Blanks or commas may separate the numbers. Note that it is not the values of the means themselves that is important, but only their differences. Thus, the mean values *0,1,2* produce the same results as the values *100,101,102*.

If too few means are entered to match the number of groups, the last mean is repeated. For example, suppose that four means are needed and you enter *1,2* (only two means). ***PASS*** will treat this as *1,2,2,2*. If too many values are entered, ***PASS*** will truncate the list to the number of means needed.

Examples:

5 20 60

2,5,7

-4,0,6,9

S Option

If an *S* is entered before the list of numbers, they are assumed to be values of σ_m, the standard deviations of the group means. A separate power calculation is made for each value. Note that this list can be a TO-BY phrase.

Examples:

S 4.7

S 4.3 5.7 4.2

S 10 to 20 by 2

Effect Size – Planned Comparisons

Contrast Coefficients

If you want to analyze a specific planned comparison, enter a set of contrast coefficients here. The calculations will then refer to the hypothesis that the corresponding contrast of the means is zero versus the alternative that it is non-zero (two-sided test). These are often called Planned Comparisons.

A contrast is a weighted average of the means in which the weights sum to zero. For example, suppose you are studying four groups and that the main hypothesis of interest is whether there is a linear trend across the groups. You would enter *-3, -1, 1, 3* here. This would form the weighted average of the means:

-3(Mean1)-(Mean2)+(Mean3)+3(Mean3)

The point to realize is that these numbers (the coefficients) are used to calculate a specific weighted average of the means which is to be compared against zero using a standard *F* (or *t*) test.

- **NONE or blank**

  When the box is left blank or the word *None* is entered, this option is ignored.

- **Linear Trend**

  A set of coefficients is generated appropriate for testing the alternative hypothesis that there is a linear (straight-line) trend across the means. These coefficients assume that the means are equally spaced across the trend variable.

- **Quadratic**

  A set of coefficients is generated appropriate for testing the alternative hypothesis that the means follow a quadratic model. These coefficients assume that the means are equally spaced across the implicit *X* variable.

- **Cubic**

  A set of coefficients is generated appropriate for testing the alternative hypothesis that the means follow a cubic model. These coefficients assume that the means are equally spaced across the implicit *X* variable.

- **First Against Others**

  A set of coefficients is generated appropriate for testing the alternative hypothesis that the first mean is different from the average of the remaining means. For example, if there were four groups, the generated coefficients would be -*3, 1, 1, 1*.

- **List of Coefficients**

  A list of coefficients, separated by commas or blanks, may be entered. If the number of items in the list does not match the number of groups (*k*), zeros are added or extra coefficients are truncated.

  Remember that these coefficients must sum to zero. Also, the scale of the coefficients does not matter. That is *0.5,0.25,0.25*; *-2,1,1*; and *-200,100,100* will yield the same results.

  To avoid rounding problems, it is better to use -*3,1,1,1* than the equivalent -*1,0.333,0.333,0.333*. The second set does not sum to zero.

Effect Size – Standard Deviation

S (Standard Deviation of Subjects)

This is σ , the standard deviation within a group. It represents the variability from subject to subject that occurs when the subjects are treated identically. It is assumed to be the same for all groups. This value is approximated in an analysis of variance table by the square root of the mean square error.

Since they are positive square roots, the numbers must be strictly greater than zero. You can press the *SD* button to obtain further help on estimating the standard deviation.

Note that if you are using this procedure to test a factor (such as an interaction) from a more complex design, the value of standard deviation is estimated by the square root of the mean square of the term that is used as the denominator in the *F* test.

You can enter a list of values separated by blanks or commas, in which case, a separate analysis will be calculated for each value.

Examples of valid entries:

1,4,7,10

1 4 7 10

1 to 10 by 3

Example 1 – Finding the Statistical Power

An experiment is being designed to compare the means of four groups using an *F* test with a significance level of either 0.01 or 0.05. Previous studies have shown that the standard deviation within a group is 18. Treatment means of 40, 10, 10, and 10 represent clinically important treatment differences. To better understand the relationship between power and sample size, the researcher wants to compute the power for several group sample sizes between 2 and 14. The sample sizes will be equal across all groups.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.01 0.05**
n (Sample Size Multiplier)	**2 to 14 by 2**
k (Number of Groups)	**4**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**40 10 10 10**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**18**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                         Std Dev    Standard
          Average           Total                        of Means   Deviation         Effect
Power           n     k         N    Alpha      Beta         (Sm)         (S)         Size
0.04238      2.00     4         8  0.01000   0.95762        12.99       18.00         0.7217
0.17513      2.00     4         8  0.05000   0.82487        12.99       18.00         0.7217
0.23886      4.00     4        16  0.01000   0.76114        12.99       18.00         0.7217
0.52165      4.00     4        16  0.05000   0.47835        12.99       18.00         0.7217
0.50581      6.00     4        24  0.01000   0.49419        12.99       18.00         0.7217
0.77327      6.00     4        24  0.05000   0.22673        12.99       18.00         0.7217
0.72695      8.00     4        32  0.01000   0.27305        12.99       18.00         0.7217
0.90642      8.00     4        32  0.05000   0.09358        12.99       18.00         0.7217
0.86702     10.00     4        40  0.01000   0.13298        12.99       18.00         0.7217
0.96514     10.00     4        40  0.05000   0.03486        12.99       18.00         0.7217
0.94143     12.00     4        48  0.01000   0.05857        12.99       18.00         0.7217
0.98802     12.00     4        48  0.05000   0.01198        12.99       18.00         0.7217
0.97623     14.00     4        56  0.01000   0.02377        12.99       18.00         0.7217
0.99614     14.00     4        56  0.05000   0.00386        12.99       18.00         0.7217

Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
n is the average group sample size.
k is the number of groups.
Total N is the total sample size of all groups.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
Sm is the standard deviation of the group means under the alternative hypothesis.
Standard deviation is the within group standard deviation.
The Effect Size is the ratio of Sm to standard deviation.

Summary Statements
In a one-way ANOVA study, sample sizes of 2, 2, 2, and 2 are obtained from the 4 groups whose
means are to be compared. The total sample of 8 subjects achieves 4% power to detect
differences among the means versus the alternative of equal means using an F test with a
0.01000 significance level. The size of the variation in the means is represented by their
standard deviation which is 12.99. The common standard deviation within a group is assumed to
be 18.00.
```

This report shows the numeric results of this power study. Following are the definitions of the columns of the report.

Power

The probability of rejecting a false null hypothesis.

Average n

The average of the group sample sizes.

k

The number of groups.

Total N

The total sample size of the study.

Alpha

The probability of rejecting a true null hypothesis. This is often called the significance level.

Beta

The probability of accepting a false null hypothesis that *Sm* is zero when *Sm* is actually equal to the value shown in the next column.

Std Dev of Means (Sm)

This is the standard deviation of the hypothesized means. It was computed from the hypothesized means. It is roughly equal to the average difference between the group means and the overall mean.

Once you have computed this, you can enter a range of values to determine the effect of the hypothesized means on the power.

Standard Deviation (S)

This is the within-group standard deviation. It was set in the Data window.

Effect Size

The effect size is the ratio of *SM* to *S*. It is an index of relative difference between the means that can be compared from study to study.

Detailed Results Report

```
Details when Alpha = 0.01000, Power = 0.04238, SM = 12.99, S = 18.00
                    Percent              Deviation          Ni
                     Ni of                   From        Times
Group         Ni   Total Ni      Mean        Mean    Deviation
1              2      25.00     40.00       22.50        45.00
2              2      25.00     10.00        7.50        15.00
3              2      25.00     10.00        7.50        15.00
4              2      25.00     10.00        7.50        15.00
ALL            8     100.00     17.50

Details when Alpha = 0.05000, Power = 0.17513, SM = 12.99, S = 18.00
                    Percent              Deviation          Ni
                     Ni of                   From        Times
Group         Ni   Total Ni      Mean        Mean    Deviation
1              2      25.00     40.00       22.50        45.00
2              2      25.00     10.00        7.50        15.00
3              2      25.00     10.00        7.50        15.00
4              2      25.00     10.00        7.50        15.00
ALL            8     100.00     17.50
```

This report shows the details of each row of the previous report.

Group

The number of the group shown on this line. The last line, labeled *ALL*, gives the average or the total as appropriate.

Ni

This is the sample size of each group. This column is especially useful when the sample sizes are unequal.

Percent Ni of Total Ni

This is the percentage of the total sample that is allocated to each group.

Mean

The is the value of the Hypothesized Mean. The final row gives the average for all groups.

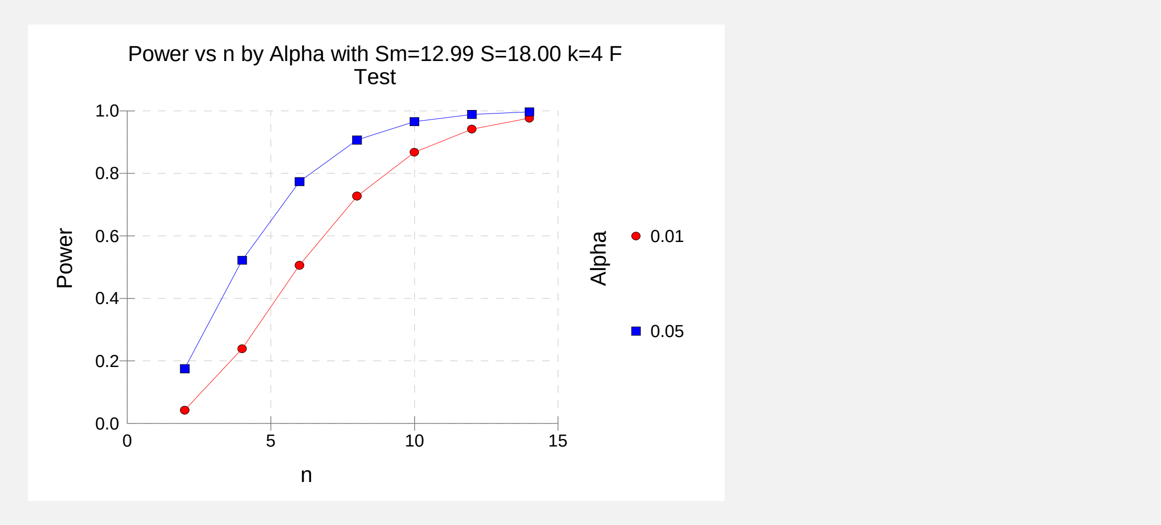
Deviation From Mean

This is the absolute value of the mean minus the overall mean. Since *Sm* is the sum of the squared deviations, these values show the relative contribution to *Sm*.

Ni Times Deviation

This is the group sample size times the absolute deviation. It shows the combined influence of the size of the deviation and the sample size on Sm.

Plots Section



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the power of increasing the sample size and increase the significance level.

When you create one of these plots, it is important to use trial and error to find an appropriate range for the horizontal variable so that you have results with both low and high power.

Example 2 – Power after a Study

This example will cover the situation in which you are calculating the power of a one-way analysis of variance *F* test on data that have already been collected and analyzed.

An experiment included a control group and two treatment groups. Each group had seven individuals. A single response was measured for each individual and recorded in the following table.

Control	T1	T2
452	646	685
674	547	658
554	774	786
447	465	536
356	759	653
654	665	669
558	767	557

When analyzed using the one-way analysis of variance procedure in ***NCSS***, the following results were obtained.

```
Analysis of Variance Table

Source                        Sum of          Mean                      Prob
Term                 DF       Squares         Square       F-Ratio      Level
A ( ... )            2        75629.8         37814.9      3.28         0.061167
S(A)                 18       207743.4        11541.3
Total (Adjusted)     20       283373.3
Total                21

Means Section

Group          Count         Mean
Control        7             527.8571
T1             7             660.4286
T2             7             649.1429
```

The significance level (Prob Level) was only 0.061—not enough for statistical significance. The researcher had hoped to show that the treatment groups had higher response levels than the control group. He could see that the group means followed this pattern since the mean for *T1* was about 25% higher than the control mean and the mean for *T2* was about 23% higher than the control mean. He decided to calculate the power of the experiment using these values of the means. (We do not recommend this approach because the power should be calculated for the minimum difference among the means that is of interest, not at the values of the sample means.)

The data entry for this problem is simple. The only entry that is not straight forward is finding an appropriate value for the standard deviation. Since the standard deviation is estimated by the square root of the mean square error, it is calculated as $\sqrt{11541.3} = 107.4304$.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**7**
k (Number of Groups)	**3**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**527.8571 660.4286 649.1429**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**107.4304**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                         Std Dev    Standard
            Average           Total                      of Means   Deviation         Effect
Power             n     k         N     Alpha       Beta     (Sm)         (S)         Size
0.54788        7.00     3        21   0.05000    0.45212    60.01      107.43         0.5586
```

The power is only 0.55. That is, there was only a 55% chance of rejecting the false null hypothesis. It is important to understand this power statement is conditional, so we will state it in detail. Given that the population means are equal to the sample means (that *Sm* is 60.01) and the population standard deviation is equal to 107.43, the probability of rejecting the false null hypothesis is 0.55. If the population means are different from the sample means (which they must be), the power is different. However, the sample means provide a reasonable place to begin.

Example 3 – Finding the Sample Size Necessary to Reject

Continuing with the last example, we will determine how large the sample size would need to have been for alpha = 0.05 and beta = 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**n (Sample Size)**
Power	**0.80**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
k (Number of Groups)	**3**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**527.8571 660.4286 649.1429**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**107.4304**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Way Analysis of Variance
                                                                Std Dev    Standard
                 Average            Total                       of Means   Deviation        Effect
Power                  n     k          N     Alpha       Beta      (Sm)         (S)        Size
0.82511            12.00     3         36   0.05000    0.17489     60.01      107.43        0.5586
```

The required sample size is 12 per group or 36 subjects.

Example 4 – Using Unequal Sample Sizes

Continuing with the last example, consider the impact of allowing the group sample sizes to be unequal. Since the control group is being compared to two treatment groups, the mean of the control group is assumed to be different from those of the treatment groups. In this situation, experience has shown that adding extra subjects to the control group can increase the power. In a separate analysis, the power with 11 subjects per group was found to be 0.7851—not quite the required 0.80.

We will try moving two subjects from each treatment group into the control group. This will give an experimental design with 15 in the control group and 9 in each of the treatment groups.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Pay particular attention to how the sample size parameters were changed. The value of n is set to one so that it is essentially ignored. The Group Sample Size Pattern contains the three unequal sample sizes.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**1**
k (Number of Groups)	**3**
Group Sample Size Pattern	**15 9 9**
Hypothesized Means	**527.8571 660.4286 649.1429**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**107.4304**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                        Std Dev     Standard
           Average           Total                      of Means    Deviation        Effect
Power            n     k         N    Alpha       Beta      (Sm)          (S)        Size
0.82967      11.00     3        33  0.05000    0.17033     63.34       107.43        0.5896
```

The power of 0.82967 achieved with the 33 subjects in this design is slightly higher than the power of 0.82511 that was achieved with the 36 subjects in the equal group size design. Apparently, unequal sample allocation can achieve better power!

We suggest that you try several different sample allocations. You will find that the optimum sample allocation depends on the values of the hypothesized means.

You should keep in mind that power may not be the only goal of the experiment. Other goals may include finding confidence intervals for each of the group means. And the narrowness of the width of the confidence interval is directly related to the sample size.

Example 5 – Minimum Detectable Difference

It may be useful to determine the minimum detectable difference among the means that can be found at the experimental conditions. This amounts to finding σ_m (which we call *Sm* on the windows and printouts).

Continuing with the previous example, find *Sm* for a wide range of sample sizes when alpha is 0.05 and beta is 0.10 or 0.20.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Sm (Std Dev of Means)**
Power	**0.80 0.90**
Alpha	**0.05**
n (Sample Size Multiplier)	**2 3 5 8 10 15 20 40 60 80 100**
k (Number of Groups)	**3**
Group Sample Size Pattern	**Equal**
Hypothesized Means	*Ignored since this is the Find setting*
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**107.4304**

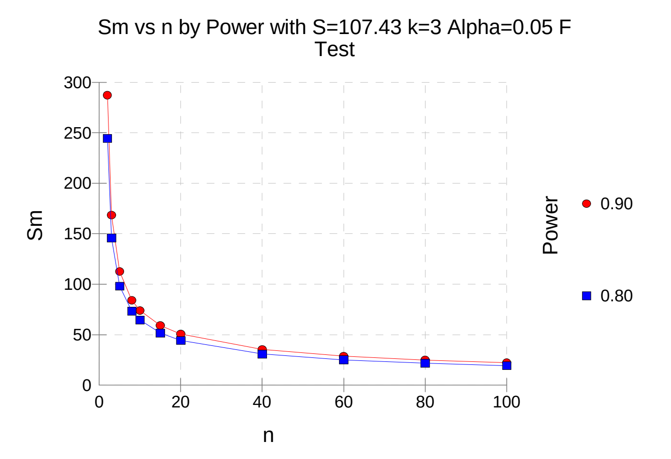
Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

Numeric Results

						Std Dev	Standard	
	Average		Total			of Means	Deviation	Effect
Power	n	k	N	Alpha	Beta	(Sm)	(S)	Size
0.90000	2.00	3	6	0.05000	0.10000	287.18	107.43	2.6732
0.80000	2.00	3	6	0.05000	0.20000	244.31	107.43	2.2741
0.90000	3.00	3	9	0.05000	0.10000	168.33	107.43	1.5669
0.80000	3.00	3	9	0.05000	0.20000	145.82	107.43	1.3573
0.90000	5.00	3	15	0.05000	0.10000	112.62	107.43	1.0483
0.80000	5.00	3	15	0.05000	0.20000	98.08	107.43	0.9130
0.90000	8.00	3	24	0.05000	0.10000	83.98	107.43	0.7817
0.80000	8.00	3	24	0.05000	0.20000	73.23	107.43	0.6817
0.90000	10.00	3	30	0.05000	0.10000	73.86	107.43	0.6875
0.80000	**10.00**	**3**	**30**	**0.05000**	**0.20000**	**64.42**	**107.43**	**0.5997**
0.90000	15.00	3	45	0.05000	0.10000	59.07	107.43	0.5499
0.80000	15.00	3	45	0.05000	0.20000	51.54	107.43	0.4797
0.90000	20.00	3	60	0.05000	0.10000	50.67	107.43	0.4716
0.80000	20.00	3	60	0.05000	0.20000	44.21	107.43	0.4115
0.90000	40.00	3	120	0.05000	0.10000	35.34	107.43	0.3289
0.80000	40.00	3	120	0.05000	0.20000	30.83	107.43	0.2870
0.90000	60.00	3	180	0.05000	0.10000	28.73	107.43	0.2674
0.80000	60.00	3	180	0.05000	0.20000	25.07	107.43	0.2333
0.90000	80.00	3	240	0.05000	0.10000	24.82	107.43	0.2311
0.80000	80.00	3	240	0.05000	0.20000	21.66	107.43	0.2016
0.90000	100.00	3	300	0.05000	0.10000	22.18	107.43	0.2064
0.80000	100.00	3	300	0.05000	0.20000	19.35	107.43	0.1801



This plot shows the relationships between power, sample size, and detectable difference. Several conclusions are possible, but the most impressive is the sharp elbow in the curve that occurs near n = 10 when *Sm* is about 64.

How do you interpret an *Sm* of 64? The easiest way is to generate a set of means that have a standard deviation of 64. To do this, press the SD button in the lower right corner of the One Way ANOVA panel to load the Standard Deviation Estimator module. Set *N* = 3, Mean = 0, and Standard Deviation = 64. Press the Two Unique Values button. This results in three means equal to -91, 45, and 45. The differences among these means are the minimum detectable differences that can be detecting with a sample size of 9 when the power is 80%.

Example 6 – Validation using Fleiss

Fleiss (1986) page 374 presents an example of determining a sample size in an experiment with 4 groups; means of 9.775, 12, 12, and 14.225; standard deviation of 3; alpha of 0.05, and beta of 0.20. He finds a sample size of 11 per group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**n (Sample Size)**
Power	**0.80**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
k (Number of Groups)	**4**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**9.775 12 12 14.225**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**3**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                           Std Dev    Standard
          Average          Total                          of Means   Deviation        Effect
Power           n     k        N     Alpha       Beta        (Sm)         (S)        Size
0.80273     11.00     4       44   0.05000    0.19727        1.57        3.00        0.5244


Details when Alpha = 0.05000, Power = 0.80273, SM = 1.57, S = 3.00
                        Percent                 Deviation          Ni
                          Ni of                      From       Times
Group          Ni      Total Ni       Mean           Mean   Deviation
1              11         25.00       9.78           2.23       24.48
2              11         25.00      12.00           0.00        0.00
3              11         25.00      12.00           0.00        0.00
4              11         25.00      14.23           2.23       24.48
ALL            44        100.00      12.00
```

PASS also found *n* = 11. Note that Fleiss used calculations based on a normal approximation, but ***PASS*** uses exact calculations based on the non-central *F* distribution.

Example 7 – Validation using Desu

Desu (1990) page 48 presents an example of determining a sample size in an experiment with 3 groups; means of 0, -0.2553, and 0.2553; standard deviation of 1; alpha of 0.05, and beta of 0.10. He finds a sample size of 99 per group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example7** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**n (Sample Size)**
Power	**0.90**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
k (Number of Groups)	**3**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**0 -0.2553 0.2553**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**1**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                              Std Dev    Standard
          Average        Total                                of Means   Deviation        Effect
Power           n    k       N    Alpha        Beta           (Sm)          (S)        Size
0.90285     99.00    3     297  0.05000     0.09715           0.21         1.00        0.2085


Details when Alpha = 0.05000, Power = 0.90285, SM = 0.21, S = 1.00
                      Percent                  Deviation           Ni
                        Ni of                       From        Times
Group            Ni  Total Ni        Mean           Mean    Deviation
1                99     33.33        0.00           0.00         0.00
2                99     33.33       -0.26           0.26        25.27
3                99     33.33        0.26           0.26        25.27
ALL             297    100.00        0.00
```

PASS also found *n* = 99.

Example 8 – Validation using Kirk

Kirk (1982) pages 140-144 presents an example of determining a sample size in an experiment with 4 groups; means of 2.75, 3.50, 6.25, and 9.0; standard deviation of 1.20995; alpha of 0.05, and beta of 0.05. He finds a sample size of 3 per group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example8** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**n (Sample Size)**
Power	**0.95**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
k (Number of Groups)	**4**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**2.75 3.5 6.25 9**
Contrast Coefficients	**None**
S (Standard Deviation of Subjects)	**1.20995**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                             Std Dev     Standard
                Average            Total                     of Means    Deviation           Effect
Power                n      k          N    Alpha       Beta     (Sm)          (S)           Size
0.99767           3.00      4         12    0.05000  0.00233     2.47         1.21           2.0376
```

PASS also found *n* = 3.

Example 9 – Power of a Planned Comparison

An experiment is being designed to study the response to different doses of a drug. Three groups, receiving a dose of 0, 10, and 20 milligrams of the drug, are anticipated. An *F* test will be used to test the hypothesis that the means exhibit a linear trend across the doses. The significance level is 0.05. Previous studies have shown the within group standard deviation to be 18. Treatment means of 5, 16, and 30 represent clinically important treatment differences. To better understand the relationship between power and sample size, the researcher wants to compute the power for several group sample sizes between 2 and 18. The sample sizes will be equal across all groups.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example9** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**2 to 18 by 2**
k (Number of Groups)	**3**
Group Sample Size Pattern	**Equal**
Hypothesized Means	**5 16 30**
Contrast Coefficients	**Linear Trend**
S (Standard Deviation of Subjects)	**18**
Axes/Legend/Grid Tab	
Vertical Range	**User (Given Below)**
Minimum	**0**
Maximum	**1**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results Report

```
Numeric Results
                                                          Std Dev    Standard
           Average          Total                        of Means   Deviation          Effect
Power            n     k        N      Alpha       Beta      (Sm)         (S)          Size
0.16781       2.00     3        6    0.05000    0.83219     10.21       18.00          0.5670
0.41889       4.00     3       12    0.05000    0.58111     10.21       18.00          0.5670
0.61410       6.00     3       18    0.05000    0.38590     10.21       18.00          0.5670
0.75458       8.00     3       24    0.05000    0.24542     10.21       18.00          0.5670
0.84932      10.00     3       30    0.05000    0.15068     10.21       18.00          0.5670
0.91013      12.00     3       36    0.05000    0.08987     10.21       18.00          0.5670
0.94768      14.00     3       42    0.05000    0.05232     10.21       18.00          0.5670
0.97017      16.00     3       48    0.05000    0.02983     10.21       18.00          0.5670
0.98329      18.00     3       54    0.05000    0.01671     10.21       18.00          0.5670

Summary Statements
In a one-way ANOVA study, sample sizes of 2, 2, and 2 are obtained from the 3 groups whose
means are to be compared using a planned comparison (contrast). The total sample of 6 subjects
achieves 17% power to detect a non-zero contrast of the means versus the alternative that the
contrast is zero using an F test with a 0.05000 significance level. The value of the contrast
of the means is 25.00. The common standard deviation within a group is assumed to be 18.00.
```

This report shows the numeric results of this power study. Most of the definitions are the same as with the one-way ANOVA test. Following are the definitions that are different.

Std Dev of Means (Sm)

When displaying results for planned comparisons, this is not the standard deviation of the hypothesized means. Instead, it is a special function of the coefficients and the hypothesized means given by the equation

$$\sigma_{mc} = \frac{\left|\sum_{i=1}^{k} c_i \mu_i\right|}{\sqrt{N \sum_{i=1}^{k} \frac{c_i^2}{n_i}}}$$

Effect Size

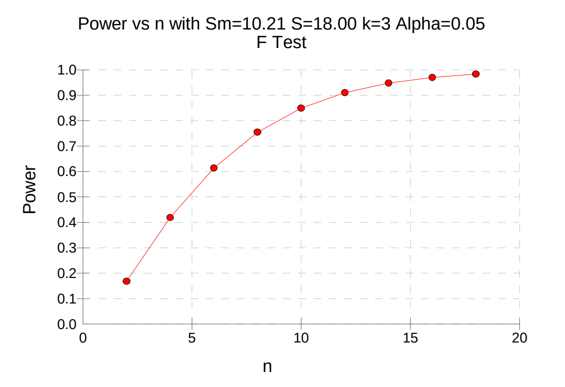
The effect size is the ratio of SM to S. It is an index of relative difference between the means that can be compared from study to study.

Details Report

```
Details when Alpha = 0.05000, Power = 0.16781, SM = 10.21, S = 18.00
                        Percent                                     Mean
                          Ni of                   Contrast         Times
Group          Ni       Total N        Mean    Coefficient      Contrast
1               2         33.33        5.00         -1.000         -5.00
2               2         33.33       16.00          0.000          0.00
3               2         33.33       30.00          1.000         30.00
ALL             6        100.00       17.00           0.00         25.00
```

This report shows the details of each row of the previous report. It is especially useful because it shows the values of the contrast coefficients and the contrast (which is the value in the lower right corner of the table).

Plots Section



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the power of increasing the sample size.

When you create one of these plots, it is important to use trial and error to find an appropriate range for the horizontal variable so that you have results with both low and high power.

Chapter 555

One-Way Analysis of Variance (Simulation)

Introduction

This procedure analyzes the power and significance level of the parametric F-Test and the nonparametric Kruskal-Wallis test which are used to test statistical hypotheses in a one-way experimental design. For each scenario that is set up, two simulations are run. One simulation estimates the significance level and the other estimates the power.

Technical Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify how the test is carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.
2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's power.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's significance level.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in ***PASS*** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draw the random numbers from this pool. This second method can cut the running time of the simulation by 70%.

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Test Statistics

Suppose *g* groups each have a normal distribution and means $\mu_1, \mu_2, \cdots, \mu_g$ and common standard deviation σ. Let $n_1, n_2, \cdots, n_g$ denote the number of subjects in each group and let *N* denote the total sample size of all groups. The tests that follow assume that the data are obtained by taking simple random samples from the *g* populations.

F-Test

The formula for the calculation of the F-test is

$$F_{g-1,N-g} = \frac{MSR}{MSE}$$

where

$$MSR = \frac{\sum_{k=1}^{g} n_k \left(\overline{X}_k - \overline{\overline{X}}\right)^2}{g-1}$$

$$MSE = \frac{\sum_{k=1}^{g} \sum_{j=1}^{n_k} \left(X_{kj} - \overline{X}_k\right)^2}{N-g}$$

$$\overline{X}_k = \frac{\sum_{j=1}^{n_k} X_{kj}}{n_k}$$

$$\overline{\overline{X}} = \frac{\sum_{k=1}^{g} n_k \overline{X}_k}{N}$$

$$N = \sum_{k=1}^{g} n_k$$

If the assumptions are met, the distribution of this test statistic follows the *F* distribution with degrees of freedom *g*-1 and *N*-*g*.

Kruskal-Wallis Test

The Kruskal-Wallis test corrected for ties is calculated using the formula

$$W = \frac{H}{T_C}$$

where

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{g} \frac{R_k^2}{n_k} - 3(N+1)$$

$$T_C = 1 - \frac{\sum t(t^2 - 1)}{N(N^2 - 1)}$$

R_k is the sum of the ranks of the kth group, and *t* is the count of a particular tie. The distribution of *W* is approximately Chi-square with *g*-1 degrees of freedom.

Procedure Options

This section describes the options that are unique to this procedure. These are located on the panels associated with the Data1, Data 2, Reports, and Options tabs. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the chapter entitled Procedure Window.

Data 1 Tab

The Data 1 tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated using the values of the other parameters. Under most conditions, you would select either *Power* or *n*.

Select *Power* when you want to estimate the power of a certain scenario.

Select *n* when you want to determine the sample size needed to achieve a given power and alpha error level. This option is very computationally intensive, so it may take a long time to complete.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

n (Sample Size Multiplier)

This is the base, per group, sample size. One or more values, separated by blanks or commas, may be entered. A separate analysis is performed for each value listed here.

The group samples sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by *m1, m2, m3, …, mk* and this value is represented by *n*, the group sample sizes *N1, N2, N3, ..., Nk* are calculated as follows:

n1=[n(m1)]

n2=[n(m2)]

n3=[n(m3)]

etc.

where the operator, [*X*] means the next integer after *X*, e.g. [3.1]=4. This is required since sample sizes must be whole numbers.

For example, suppose there are three groups and the Group Sample Size Pattern is set to *1,2,3*. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to *2,4,6,8,10*; five sets of group sample sizes will be generated and an analysis run for each. These sets are:

2	4	6
4	8	12
6	12	18
8	16	24
10	20	30

As a second example, suppose there are three groups and the Group Sample Size Pattern is *0.2,0.3,0.5*. When the fractional Pattern values sum to one, *n* can be interpreted as the total sample size *N* of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

(0.2)12 = 2.4 which is rounded up to the next whole integer, 3.

(0.3)12 = 3.6 which is rounded up to the next whole integer, 4.

(0.5)12 = 6.

Note that in this case, 3+4+6 does not equal n (which is 12). This can happen because of rounding.

Group Sample Size Pattern

A set of positive, numeric values, one for each row of distributions, is entered here. Each item specified in this list applies to the whole row of distributions. For example, suppose the entry is *1 2 1* and Grps 1 = 3, Grps 2 = 1, Grps 3 = 2. The sample size pattern used would be *1 1 1 2 1 1*.

The sample size of group *i* is found by multiplying the ith number from this list by the value of *n* and rounding up to the next whole number. The number of values must match the number of groups, *g*. When too few numbers are entered, 1's are added. When too many numbers are entered, the extras are ignored.

- **Equal**

  If all sample sizes are to be equal, enter *Equal* here and the desired sample size in *n*. A set of *g* 1's will be used. This will result in *n1* = *n2* = … = *ng* = *n*. That is, all sample sizes are equal to *n*.

Test

Test Statistic

Specify which test is to be simulated. Although the F-test is the most commonly used test, it is based on assumptions that may not be viable in some situations. For your data, you may find that the Kruskal-Wallis test is more accurate (actual alpha = target alpha) and more precise (better power).

Simulations

Simulations

This option specifies the number of iterations, *M*, used in the simulation. The larger the number of iterations, the longer the running time, and, the more accurate the results.

The precision of the simulated power estimates are calculated from the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the 'Precision' amount given in the table.

Simulation	Precision	Precision
Size	when	when
M	Power = 0.50	Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

These options specify the distributions to be used in the two simulations, one set per row. The first option specifies the number of groups represented by the two distributions that follow. The second option specifies the distribution to be used in simulating the null hypothesis to determine the significance level (alpha). The third option specifies the distribution to be used in simulating the alternative hypothesis to determine the power.

Grps [1 – 3] (Grps 4 – 9 are found on the Data 2 tab)

This value specifies the number of groups specified by the H0 and H1 distribution statements to the right. Usually, you will enter '1' to specify a single H0 and a single H1 distribution, or you will enter '0' to indicate that the distributions specified on this line are to be ignored. This option lets you easily specify many identical distributions with a single phrase.

The total number of groups *g* is equal to the sum of the values for the three rows of distributions shown under the Data1 tab and the six rows of distributions shown under the Data2 tab.

Note that each item specified in the 'Group Sample Size Pattern' option applies to the whole row of entries here. For example, suppose the 'Group Sample Size Pattern' was '1 2 1' and 'Grps 1' = 3, 'Grps 2' = 1, and 'Grps 3' = 2. The sample size pattern would be '1 1 1 2 1 1'.

Group Distribution(s)|H0

This entry specifies the distribution of one or more groups under the null hypothesis, H0. The magnitude of the differences of the means of these distributions, which is often summarized as the standard deviation of the means, represents the magnitude of the mean differences specified under H0. Usually, the means are assumed to be equal under H0, so their standard deviation should be zero except for rounding.

These distributions are used in the simulations that estimate the actual significance level. They also specify the value of the mean under the null hypothesis, H0. Usually, these distributions will be identical. The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean is entered first.

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means, since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Group Distribution(s)|H1

Specify the distribution of this group under the alternative hypothesis, H1. This distribution is used in the simulation that determines the power. A fundamental quantity in a power analysis is the amount of variation among the group means. In fact, classical power analysis formulas, this variation is summarized as the standard deviation of the means.

The important point to realize is that you must pay particular attention to the values you give to the means of these distributions because they are fundamental to the interpretation of the simulation.

For convenience in specifying a range of values, the parameters of the distribution can be specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean, *M1*, is entered first.

```
Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)
```

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for *M0* in the distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for *M1* in the distribution specifications given above. Although it can be used wherever you want, *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B, C)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values for each letter using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Reports Tab

The Reports tab contains settings about the format of the output.

Select Output – Numeric Reports

Show Numeric Reports & Plots

These options let you specify whether you want to generate the standard reports and plots.

Show Inc's & 95% C.I.

Checking this option causes an additional line to be printed showing a 95% confidence interval for both the power and actual alpha and half the width of the confidence interval (the increment).

Select Output – Plots

Show Comparative Reports & Plots

These options let you specify whether you want to generate reports and plots that compare the test statistics that are available.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Random Numbers

Random Number Pool Size

This is the size of the pool of values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Example 1 – Power at Various Sample Sizes

For this first example we repeat Example 1 of the regular One-Way ANOVA procedure. This will allow you to compare the values obtained by simulation with the actual values obtained from the theoretical results.

An experiment is being designed to compare the means of four groups using an *F* test with a significance level of 0.05. Previous studies have shown that the standard deviation within a group is 18. Treatment means of 40, 10, 10, and 10 represent clinically important treatment differences. To better understand the relationship between power and sample size, the researcher wants to compute the power for group sample sizes of 4, 8, and 12. The group sample sizes are equal.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance (Simulation)** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**4 8 12**
Group Sample Size Pattern	**Equal**
Test Type	**F-Test**
Simulations	**2000**
Grps 1	**1**
Group 1 Distribution(s) | H0	**N(M0 S)**
Group 1 Distribution(s) | H1	**N(M1 S)**
Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M0 S)**
Grps 3	**0**
M0 (Mean|H0)	**10**
M1 (Mean|H1)	**40**
S	**18**
Options Tab	
Random Number Pool Size	**Automatic**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results Report

```
Numeric Results for Testing the g = 4 Group Means
Test Statistic: F-Test

          Average    Total
            Group   Sample                                    S.D. of   S.D. of
             Size     Size    Target    Actual                  Means      Data
Power           n        N     Alpha     Alpha       Beta       Sm|H1     SD|H1        M0         M1          S
0.536         4.0       16     0.050     0.047      0.465        13.0      18.0      10.0       40.0       18.0
(0.022)    [0.514   0.557]              (0.009)    [0.038      0.056]

0.896         8.0       32     0.050     0.053      0.105        13.0      18.1      10.0       40.0       18.0
(0.013)    [0.882   0.909]              (0.010)    [0.043      0.063]

0.982        12.0       48     0.050     0.051      0.019        13.0      18.1      10.0       40.0       18.0
(0.006)    [0.976   0.987]              (0.010)    [0.041      0.061]

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 12.84 seconds.
H0 Distributions: Normal(M0 S); Normal(M0 S); Normal(M0 S); and Normal(M0 S)
H1 Distributions: Normal(M1 S); Normal(M0 S); Normal(M0 S); and Normal(M0 S)

Report Definitions
H0 represents the null hypothesis.
H1 represents the alternative hypothesis.
Power is the probability of rejecting a false null hypothesis. It should be close to one.
'n' is the average of the group sample sizes.
Total N is the total sample size of all groups combined.
Target Alpha is the desired probability of rejecting a true null hypothesis.
Actual Alpha is the alpha achieved by this simulation.
Beta is the probability of accepting a false null hypothesis.
Sm|H1 is the standard deviation of the group means under the alternative hypothesis.
SD|H1 is the within group standard deviation under the alternative hypothesis.
Second Row: (Power Prec.) [95% LCL and UCL Power]    (Alpha Prec.) [95% LCL and UCL Alpha]

Summary Statements
A one-way design with 4 groups has sample sizes of 4, 4, 4, and 4. The null hypothesis is that
the standard deviation of the group means is 0.1 and the alternative standard deviation of the
group means is 13.0. The total sample of 16 subjects achieves a power of 0.536 using the F-Test
with a target significance level of 0.050 and an actual significance level of 0.047. The
average within group standard deviation assuming the alternative distribution is 18.0. These
results are based on 2000 Monte Carlo samples from the null distributions: Normal(M0 S);
Normal(M0 S); Normal(M0 S); and Normal(M0 S) and the alternative distributions: Normal(M1 S);
Normal(M0 S); Normal(M0 S); and Normal(M0 S). Other parameters used in the simulation were: M0
= 10.0, M1 = 40.0, and S = 18.0.
```

This report shows the estimated power for each scenario. The first row shows the parameter settings and the estimated power and significance level (Actual Alpha).

The second row shows two 95% confidence intervals in brackets: the first for the power and the second for the significance level. Half the width of each confidence interval is given in parentheses as a fundamental measure of the accuracy of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Power

This is the probability of rejecting a false null hypothesis. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values are provided to help you understand the precision of the estimated power.

Average Group Size n

This is the average of the group sample sizes.

Total Sample Size N

This is the total sample size of the study.

Target Alpha

The target value of alpha: the probability of rejecting a true null hypothesis. This is often called the significance level.

Actual Alpha

This is the value of alpha estimated by the simulation using the H0 distributions. It should be compared with the Target Alpha to determine if the test statistic is accurate in this scenario.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values are provided to help you understand the precision of the Actual Alpha.

Beta

Beta is the probability of accepting a false null hypothesis. This is the value of beta estimated by the simulation using the H1 distributions.

S.D. of Means Sm|H1

This is the standard deviation of the hypothesized means of the alternative distributions. Under the null hypothesis, this value is zero. So this value represents the magnitude of the difference among the means that is being tested. It is roughly equal to the average difference between the group means and the overall mean.

Note that the effect size is the ratio of Sm|H1 and SD|H1.

S.D. of Data SD|H1

This is the within-group standard deviation calculated from samples from the alternative distributions.

M0

This is the value entered for M0, the group means under H0.

M1

This is the value entered for M1, the group means under H1.

S

This is the value entered for S, the standard deviation.

Detailed Results Report

```
Details when Target Alpha = 0.050, M0 = 10.0, M1 = 40.0, S = 18.0
Test Statistic: F-Test
                   Percent      H0       H1       H0       H1       H0       H1   Actual
Group          Ni  Ni of N    Mean     Mean     S.D.     S.D.       Sm       Sm    Alpha    Power
All            16   100.00    10.0     17.5     18.1     18.0      0.0     13.0    0.051    0.536
1               4    25.00    10.0     40.0     17.7     17.9
2               4    25.00    10.0     10.0     18.4     18.2
3               4    25.00    10.0     10.0     18.2     17.8
4               4    25.00    10.0     10.0     18.0     18.1

Details when Target Alpha = 0.050, M0 = 10.0, M1 = 40.0, S = 18.0
Test Statistic: F-Test
                   Percent      H0       H1       H0       H1       H0       H1   Actual
Group          Ni  Ni of N    Mean     Mean     S.D.     S.D.       Sm       Sm    Alpha    Power
All            32   100.00    10.0     17.5     18.0     18.1      0.0     13.0    0.051    0.896
1               8    25.00    10.0     40.0     17.9     18.1
2               8    25.00    10.0     10.0     18.1     18.4
3               8    25.00    10.0     10.0     18.1     18.1
4               8    25.00    10.0     10.0     18.0     18.0

Details when Target Alpha = 0.050, M0 = 10.0, M1 = 40.0, S = 18.0
Test Statistic: F-Test
                   Percent      H0       H1       H0       H1       H0       H1   Actual
Group          Ni  Ni of N    Mean     Mean     S.D.     S.D.       Sm       Sm    Alpha    Power
All            48   100.00    10.0     17.5     18.1     18.1      0.0     13.0    0.051    0.982
1              12    25.00    10.0     40.0     18.3     18.0
2              12    25.00    10.0     10.0     18.1     18.4
3              12    25.00    10.0     10.0     18.2     18.0
4              12    25.00    10.0     10.0     17.9     17.9
```

This report shows the details of each row of the previous report.

Group

This is the number of the group shown on this line. The first line, labeled *All*, gives the average or the total as appropriate.

Ni

This is the sample size of each group. This column is especially useful when the sample sizes are unequal.

Percent Ni of N

This is the percentage of the total sample that is allocated to each group.

H0 and H1 Means

These are the means that were used in the simulations for H0 and H1, respectively.

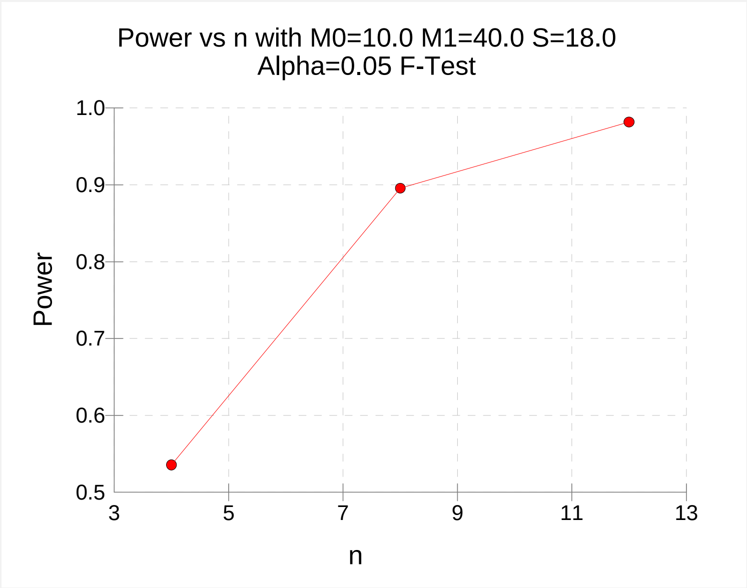
H0 and H1 S.D.'s

These are the standard deviations that were obtained by the simulations for H0 and H1, respectively. Note that they often are not exactly equal to what was specified because of simulation error.

H0 and H1 Sm's

These are the standard deviations of the means that were obtained by the simulations for H0 and H1, respectively. Under H0, the value of Sm should be near zero. It lets you determine if your simulation of H0 was correctly specified.

Plots Section



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the power of increasing the sample size.

Example 2 – Comparative Results

Continuing with Example 1, the researchers want to study the characteristics of alternative test statistics. They want to compare the results of the F-test and the Kruskal-Wallis test.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance (Simulation)** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**4 8 12**
Group Sample Size Pattern	**Equal**
Test Type	**F-Test**
Simulations	**2000**
Grps 1	**1**
Group 1 Distribution(s) | H0	**N(M0 S)**
Group 1 Distribution(s) | H1	**N(M1 S)**

Data Tab (continued)

Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M0 S)**
Grps 3	**0**
M0 (Mean|H0)	**10**
M1 (Mean|H1)	**40**
S	**18**

Reports Tab

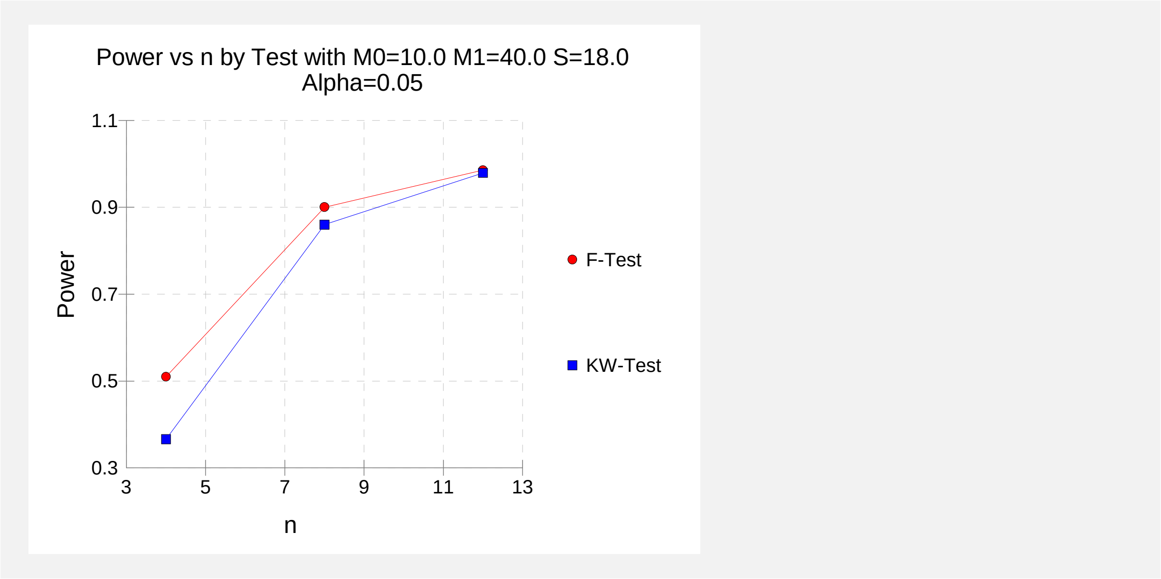
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Numeric Results for Testing the g = 4 Group Means
Test Statistic: F-Test

          Average   Total
            Group  Sample                              S.D. of   S.D. of
             Size    Size   Target   Actual            Means      Data
Power           n       N    Alpha    Alpha     Beta   Sm|H1     SD|H1        M0        M1         S
0.510         4.0      16    0.050    0.051    0.490    13.0      17.8      10.0      40.0      18.0
0.901         8.0      32    0.050    0.047    0.100    13.0      18.0      10.0      40.0      18.0
0.986        12.0      48    0.050    0.046    0.015    13.0      18.1      10.0      40.0      18.0

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 21.02 seconds.
H0 Distributions: Normal(M0 S); Normal(M0 S); Normal(M0 S); and Normal(M0 S)
H1 Distributions: Normal(M1 S); Normal(M0 S); Normal(M0 S); and Normal(M0 S)

Power Comparison for Testing the g = 4 Group Means

Total     S.D. of  S.D. of                    Kruskal
Sample    Means       Data   Target   F-Test    Wallis
Size      Sm|H1      SD|H1    Alpha    Power     Power        M0        M1         S
16           13.0     17.8    0.050    0.510     0.366      10.0      40.0      18.0
32           13.0     18.0    0.050    0.901     0.860      10.0      40.0      18.0
48           13.0     18.1    0.050    0.986     0.979      10.0      40.0      18.0

Alpha Comparison for Testing the g = 4 Group Means

Total     S.D. of  S.D. of                    Kruskal
Sample    Means       Data   Target   F-Test    Wallis
Size      Sm|H1      SD|H1    Alpha    Alpha     Alpha        M0        M1         S
16           13.0     16.5    0.050    0.042     0.026      10.0      40.0      18.0
16           13.0     17.8    0.050    0.051     0.041      10.0      40.0      18.0
32           13.0     18.0    0.050    0.047     0.042      10.0      40.0      18.0
48           13.0     18.1    0.050    0.046     0.039      10.0      40.0      18.0
```



We notice that the power of the F-test is much greater than the Kruskal-Wallis test for *n* = 4. However, when *n* = 12, the powers of the two tests are almost equal. Note that the alpha value of the Kruskal-Wallis test is almost half that of the F-test for *n* = 4. This is probably why the power is also low.

Example 3 – Validation using Fleiss

Fleiss (1986) page 374 presents an example of determining an appropriate sample size when using an F-test in an experiment with 4 groups; means of 9.775, 12, 12, and 14.225; standard deviation of 3; alpha of 0.05, and beta of 0.20. He finds a sample size of 11 per group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance (Simulation)** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	**0.80**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
Group Sample Size Pattern	**Equal**
Test Type	**F-Test**
Simulations	**2000**
Grps 1	**1**
Group 1 Distribution(s) | H0	**N(M0 S)**
Group 1 Distribution(s) | H1	**N(9.775 S)**

Data Tab (continued)

Grps 2	**2**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M0 S)**
Grps 3	**1**
Group 3 Distribution(s) | H0	**N(M0 S)**
Group 3 Distribution(s) | H1	**N(14.225 S)**
M0 (Mean|H0)	**12**
M1 (Mean|H1)	**0**
S	**3**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing the g = 4 Group Means
Test Statistic: F-Test

           Average    Total
             Group   Sample                                 S.D. of   S.D. of
              Size     Size   Target    Actual               Means      Data
Power            n        N    Alpha     Alpha      Beta     Sm|H1     SD|H1        M0         S
0.822         11.0       44    0.050     0.052     0.179       1.6       3.0      12.0       3.0
(0.017)     [0.805    0.838]           (0.010)    [0.042    0.061]

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 18.88 seconds.
H0 Distributions: Normal(M0 S); Normal(M0 S); Normal(M0 S); and Normal(M0 S)
H1 Distributions: Normal(9.775 S); Normal(M0 S); Normal(M0 S); and Normal(14.225 S)

Details when Target Alpha = 0.050, M0 = 12.0, S = 3.0
Test Statistic: F-Test
                         Percent       H0        H1        H0        H1        H0        H1    Actual
Group               Ni   Ni of N     Mean      Mean      S.D.      S.D.        Sm        Sm     Alpha     Power
All                 44    100.00     12.0      12.0       3.0       3.0       0.0       1.6     0.052     0.822
1                   11     25.00     12.0       9.8       3.0       3.0
2                   11     25.00     12.0      12.0       3.0       3.0
3                   11     25.00     12.0      12.0       3.0       3.0
4                   11     25.00     12.0      14.2       3.0       3.0
```

Note that ***PASS*** has also found the group sample size to be 11.

Example 4 – Selecting a Test Statistic when the Data Contain Outliers

The F-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracy when the data contain outliers. This example will investigate the impact of outliers on the power and precision of the F-test and the Kruskal-Wallis test.

A mixture of two normal distributions will be used to randomly generate outliers. The mixture will draw 95% of the data from a normal distribution with mean zero and variance one. The other 5% of the data will come from a normal distribution with mean zero and variance that ranges from one to ten. In the alternative distributions, two will have a mean of zero and one will have a mean of one.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance (Simulation)** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

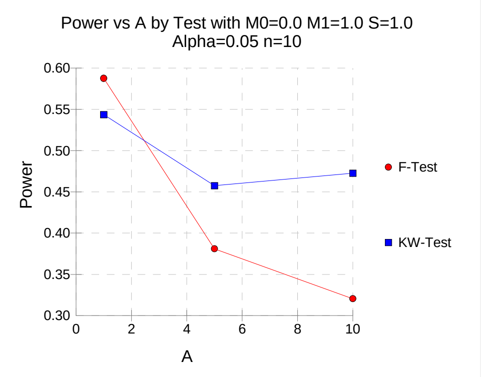
Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**10 20**
Group Sample Size Pattern	**Equal**
Test Type	**F-Test**
Simulations	**2000**
Grps 1	**2**
Group 1 Distribution(s) | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 1 Distribution(s) | H1	**N(M0 S)[95];N(M0 A)[5]**
Grps 2	**1**
Group 2 Distribution(s) | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Distribution(s) | H1	**N(M1 S)[95];N(M1 A)[5]**
Grps 3	**0**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**1**
A	**1 5 10**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

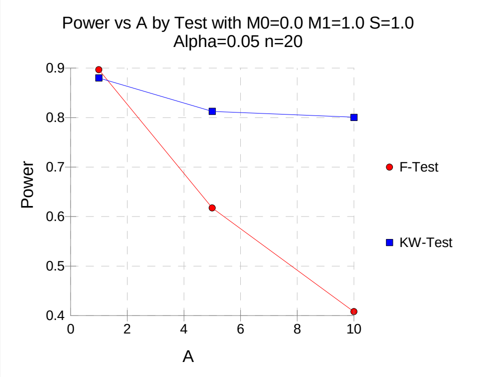
Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Testing the g = 3 Group Means
Test Statistic: F-Test

        Average     Total
          Group    Sample                                S.D. of   S.D. of
           Size      Size    Target    Actual              Means      Data
Power         n         N     Alpha     Alpha      Beta    Sm|H1     SD|H1       M0      M1        S        A
0.588      10.0        30     0.050     0.054     0.413      0.5       1.0      0.0     1.0      1.0      1.0
0.381      10.0        30     0.050     0.035     0.619      0.5       1.4      0.0     1.0      1.0      5.0
0.321      10.0        30     0.050     0.026     0.680      0.5       2.4      0.0     1.0      1.0     10.0
0.897      20.0        60     0.050     0.049     0.104      0.5       1.0      0.0     1.0      1.0      1.0
0.618      20.0        60     0.050     0.047     0.383      0.5       1.5      0.0     1.0      1.0      5.0
0.408      20.0        60     0.050     0.026     0.592      0.5       2.4      0.0     1.0      1.0     10.0

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 40.30 seconds.
H0 Distributions: Normal(M0 S)[95];Normal(M0 A)[5]; Normal(M0 S)[95];Normal(M0 A)[5]; and Normal(M0
S)[95];Normal(M0 A)[5]
H1 Distributions: Normal(M1 S)[95];Normal(M1 A)[5]; Normal(M0 S)[95];Normal(M0 A)[5]; and Normal(M0
S)[95];Normal(M0 A)[5]

Power Comparison for Testing the g = 3 Group Means

Total     S.D. of   S.D. of                     Kruskal
Sample      Means      Data    Target    F-Test   Wallis
Size        Sm|H1     SD|H1     Alpha     Power    Power        M0          M1           S          A
30            0.5       1.0     0.050     0.588    0.544       0.0         1.0         1.0        1.0
30            0.5       1.4     0.050     0.381    0.458       0.0         1.0         1.0        5.0
30            0.5       2.4     0.050     0.321    0.473       0.0         1.0         1.0       10.0
60            0.5       1.0     0.050     0.897    0.880       0.0         1.0         1.0        1.0
60            0.5       1.5     0.050     0.618    0.813       0.0         1.0         1.0        5.0
60            0.5       2.4     0.050     0.408    0.801       0.0         1.0         1.0       10.0

Alpha Comparison for Testing the g = 3 Group Means

Total     S.D. of   S.D. of                     Kruskal
Sample      Means      Data    Target    F-Test   Wallis
Size        Sm|H1     SD|H1     Alpha     Alpha    Alpha        M0          M1           S          A
30            0.5       1.0     0.050     0.054    0.055       0.0         1.0         1.0        1.0
30            0.5       1.4     0.050     0.035    0.041       0.0         1.0         1.0        5.0
30            0.5       2.4     0.050     0.026    0.041       0.0         1.0         1.0       10.0
60            0.5       1.0     0.050     0.049    0.050       0.0         1.0         1.0        1.0
60            0.5       1.5     0.050     0.047    0.054       0.0         1.0         1.0        5.0
60            0.5       2.4     0.050     0.026    0.048       0.0         1.0         1.0       10.0
```


Power vs A by Test with M0=0.0 M1=1.0 S=1.0
Alpha=0.05 n=10


Power vs A by Test with M0=0.0 M1=1.0 S=1.0
Alpha=0.05 n=20

We note that when the variances are equal (A = 1), the F-Test is slightly better than the Kruskal-Wallis test. However, as the number of outliers is increased, the F-test does increasingly worse both in terms of power and significance, but the Kruskal-Wallis test is considerably less affected.

Example 5 – Selecting a Test Statistic when the Data are Skewed

The F-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracy when the underlying distributions are skewed. This example will investigate the impact of skewness on the power and precision of the F-test and the Kruskal-Wallis test.

Tukey’s lambda distribution will be used because it allows the amount of skewness to be gradually increased.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Way Analysis of Variance (Simulation)** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **One-Way ANOVA (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

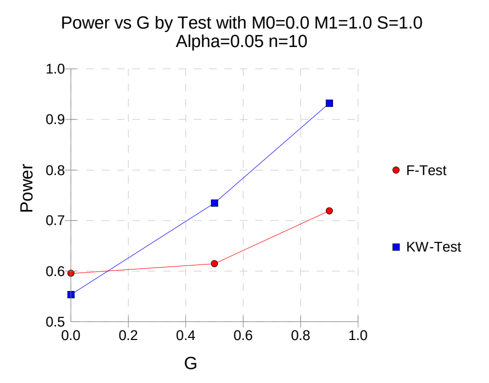
Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**10 20**
Group Sample Size Pattern	**Equal**
Test Type	**F-Test**
Simulations	**2000**
Grps 1	**2**
Group 1 Distribution(s) | H0	**L(M0 S G 0)**
Group 1 Distribution(s) | H1	**L(M0 S G 0)**
Grps 2	**1**
Group 2 Distribution(s) | H0	**L(M0 S G 0)**
Group 2 Distribution(s) | H1	**L(M1 S G 0)**
Grps 3	**0**
M0 (Mean|H0)	**0**
M1 (Mean|H0)	**1**
S	**1**
G	**0 0.5 0.9**
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

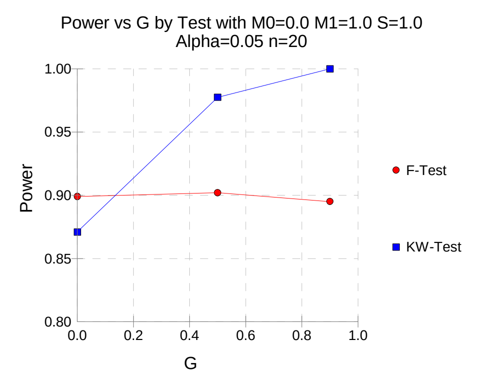
Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Numeric Results for Testing the g = 3 Group Means
Test Statistic: F-Test

        Average      Total
          Group     Sample                                   S.D. of   S.D. of
           Size       Size   Target   Actual                   Means      Data
Power         n          N    Alpha    Alpha     Beta        Sm|H1     SD|H1      M0    M1      S      G
0.596      10.0         30    0.050    0.050    0.405          0.5       1.0     0.0   1.0    1.0    0.0
0.615      10.0         30    0.050    0.034    0.386          0.5       1.0     0.0   1.0    1.0    0.5
0.719      10.0         30    0.050    0.036    0.281          0.5       1.0     0.0   1.0    1.0    0.9
0.899      20.0         60    0.050    0.046    0.101          0.5       1.0     0.0   1.0    1.0    0.0
0.902      20.0         60    0.050    0.050    0.098          0.5       1.0     0.0   1.0    1.0    0.5
0.895      20.0         60    0.050    0.034    0.105          0.5       1.0     0.0   1.0    1.0    0.9

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 40.30 seconds.
H0 Distributions: Tukey(M0 S G 0); Tukey(M0 S G 0); and Tukey(M0 S G 0)
H1 Distributions: Tukey(M0 S G 0); Tukey(M0 S G 0); and Tukey(M1 S G 0)

Power Comparison for Testing the g = 3 Group Means

Total    S.D. of  S.D. of                      Kruskal
Sample     Means     Data   Target   F-Test     Wallis
Size       Sm|H1    SD|H1    Alpha    Power      Power      M0        M1         S         G
30           0.5      1.0    0.050    0.596      0.554     0.0       1.0       1.0       0.0
30           0.5      1.0    0.050    0.615      0.735     0.0       1.0       1.0       0.5
30           0.5      1.0    0.050    0.719      0.932     0.0       1.0       1.0       0.9
60           0.5      1.0    0.050    0.899      0.871     0.0       1.0       1.0       0.0
60           0.5      1.0    0.050    0.902      0.978     0.0       1.0       1.0       0.5
60           0.5      1.0    0.050    0.895      1.000     0.0       1.0       1.0       0.9

Alpha Comparison for Testing the g = 3 Group Means

Total    S.D. of  S.D. of                      Kruskal
Sample     Means     Data   Target   F-Test     Wallis
Size       Sm|H1    SD|H1    Alpha    Alpha      Alpha      M0        M1         S         G
30           0.5      1.0    0.050    0.050      0.045     0.0       1.0       1.0       0.0
30           0.5      1.0    0.050    0.034      0.039     0.0       1.0       1.0       0.5
30           0.5      1.0    0.050    0.036      0.047     0.0       1.0       1.0       0.9
60           0.5      1.0    0.050    0.046      0.041     0.0       1.0       1.0       0.0
60           0.5      1.0    0.050    0.050      0.046     0.0       1.0       1.0       0.5
60           0.5      1.0    0.050    0.034      0.055     0.0       1.0       1.0       0.9
```





We note that as the skewness increases, the power of the Kruskal-Wallis test increases substantially as compared to the F-test.

Chapter 560

Fixed Effects Analysis of Variance

Introduction

A common task in research is to compare the average response across levels of one or more factor variables. Examples of factor variables are income level of two regions, nitrogen content of three lakes, or drug dosage. The fixed-effects analysis of variance compares the means of two or more factors. *F* tests are used to determine statistical significance of the factors and their interactions. The tests are nondirectional in that the null hypothesis specifies that all means are equal and the alternative hypothesis simply states that at least one mean is different. This ***PASS*** module performs power analysis and sample size estimation for an analysis of variance design with up to three factors.

In the following example, the responses of a weight loss experiment are arranged in a two-factor, fixed-effect, design. The first factor is diet (D1 and D2) and the second factor is dose level of a dietary drug (low, medium, and high). The twelve individuals available for this study were assigned at random to one of the six treatment groups (cells) so that there were two per group. The response variable was an individual's weight loss after four months.

Table of Individual Weight Losses

	Dietary Drug Dose Level		
Diet	Low	Medium	High
D1	14, 16	15, 18	23, 28
D2	18, 21	18, 22	38, 39

Important features to note are that each table entry represents a different individual and that the response variable (weight loss) is continuous, while the factors (Diet and Dose) are discrete.

Means can be calculated for each cell of the table. These means are shown in the table below. Note that we have added an additional row and column for the row, column, and overall means. The six means in the interior of this table are called the *cell means*.

Table of Means

	Dietary Drug Dose Level			
Diet	**Low**	**Medium**	**High**	**Total**
D1	15.00	16.50	25.50	**19.00**
D2	19.50	20.00	38.50	**26.00**
Total	**17.25**	**18.25**	**32.00**	**22.50**

The Linear Model

A mathematical model may be formulated that underlies this experimental design. This model expresses each cell mean, μ_{ij}, as the sum of parameters called *effects*. A common linear model for a two-factor experiment is

$$\mu_{ij} = m + a_i + b_j + (ab)_{ij}$$

where *i* = 1, 2, ..., *I* and *j* = 1, 2, ..., *J*. This model expresses the value of a cell mean as the sum of four components:

m the grand mean.

a_i the effect of the i^{th} level of factor *A*. Note that $\sum a_i = 0$.

b_j the effect of the j^{th} level of factor *B*. Note that $\sum b_j = 0$.

ab_{ij} the combined effect of the i^{th} level of factor A and the j^{th} level of factor *B*. Note that $\sum (ab)_{ij} = 0$.

Another way of stating this model for the two factor case is

Cell Mean = *Overall Effect* + *Row Effect* + *Column Effect* + *Interaction Effect*.

Since this model is the sum of various constants, it is called a *linear model*.

Calculating the Effects

We will now calculate the effects for our example. We will let Drug Dose correspond to factor A and Diet correspond to factor B.

Step 1 – Remove the Grand Mean

Remove the grand mean from the table of means by subtracting 22.50 from each entry. The values in the margins are the *effects* of the corresponding factors.

Table of Mean Weight Losses After Subtracting the Grand Mean

	Dietary Drug Dose Level			
Diet	**Low**	**Medium**	**High**	**Overall**
D1	-7.50	-6.00	3.00	-3.50
D2	-3.00	-2.50	16.00	3.50
Overall	-5.25	-4.25	9.50	22.50

Step 2 – Remove the Effects of Factor B (Diet)

Subtract the Diet effects (-3.50 and 3.50) from the entries in those rows.

Table of Mean Weight Losses After Subtracting the Diet Effects

	Dietary Drug Dose Level			
Diet	**Low**	**Medium**	**High**	**Overall**
D1	-4.00	-2.50	6.50	-3.50
D2	-6.50	-6.00	12.50	3.50
Overall	-5.25	-4.25	9.50	22.50

Step 3 – Remove the Effects of Factor A (Drug Dose)

Subtract the Drug Dose effects (-5.25, -4.25, and 9.50) from the rest of the entries in those columns. This will result in a table of effects.

Table of Effects

	Dietary Drug Dose Level			
Diet	**Low**	**Medium**	**High**	**Overall**
D1	1.25	1.75	-3.00	-3.50
D2	-1.25	-1.75	3.00	3.50
Overall	-5.25	-4.25	9.50	22.50

We have calculated a table of effects for the two-way linear model. Each cell mean can calculated by summing the appropriate entries from this table.

The estimated linear effects are:

m = 22.50		
a1 = -5.25	a2 = -4.25	a3 = 9.50
b1 = -3.50	b2 = 3.50	
ab11 = 1.25	ab21 = 1.75	ab31 = -3.00
ab12 = -1.25	ab22 = -1.75	ab32 = 3.00.

The six cell means are calculated from these effects as follows:

15.00 = 22.50 - 5.25 - 3.50 + 1.25

19.50 = 22.50 - 5.25 + 3.50 - 1.25

16.50 = 22.50 - 4.25 - 3.50 + 1.75

20.00 = 22.50 - 4.25 + 3.50 - 1.75

25.50 = 22.50 + 9.50 - 3.50 - 3.00

38.50 = 22.50 + 9.50 + 3.50 + 3.00

Analysis of Variance Hypotheses

The hypotheses that are tested in an analysis of variance table concern the effects, so in order to conduct a power analysis you must have a firm grasp of their meaning. For example, we would usually test the following hypotheses:

1. Are there differences in weight loss among the three drug doses? That is, are the drug dose effects all zero? This hypothesis is tested by the *F* test for factor *A*, which tests whether the standard deviation of the a_i is zero.
2. Is there a difference in weight loss between the two diets? That is, are the diet effects all zero? This hypothesis is tested by the *F* test for factor *B*, which tests whether the standard deviation of the b_j is zero.
3. Are there any diet-dose combinations that exhibit a weight loss that cannot be explained by diet and/or drug dose singly? This hypothesis is tested by the *F* test for the *AB* interaction, which tests whether the standard deviation of the $(ab)_{ij}$ is zero.

Each of these hypotheses can be tested at a different alpha level and different precision. Hence each can have a different power. One of the tasks in planning such an experiment is to determine a sample size that yields necessary power values for each of these hypothesis tests. This is accomplished using this program module.

Definition of Terms

Factorial designs evaluate the effect of two or more categorical variables (called *factors*) on a response variable by testing hypotheses about various averages. These designs are popular because they allow experimentation across a wide variety of conditions and because they evaluate the *interaction* of two or more factors. Interaction is the effect that may be attributed to a combination of two or more factors, but not to one factor singly.

A *factor* is a variable that relates to the response. Either the factor is discrete by nature (as in location or gender) or has been made discrete by collapsing a continuous variable (as in income level or age group). The term *factorial* implies that all possible combinations of the factors being studied are included in the design.

A *fixed* factor is one in which all possible *levels* (categories) are considered. Examples of fixed factors are gender, dose level, and country of origin. They are different from *random* factors which represent a random selection of individuals from the population described by the factor. Examples of random factors are people living within a region, a sample of schools in a state, or a selection of labs. Again, a fixed factor includes the range of interest while a random factor includes only a sample of all possible levels.

A factorial design is analyzed using the analysis of variance. When only fixed factors are used in the design, the analysis is said to be a *fixed-effects analysis of variance*. Other types of designs will be discussed in later chapters.

Suppose a group of individuals have agreed to be in a study involving six treatments. In a *completely randomized factorial design*, each individual is assigned at random to one of the six groups and then the treatments are applied. In some situations, the randomization occurs by randomly selecting individuals from the populations defined by the treatment groups. The designs analyzed by this module are completely randomized factorial designs.

Power Calculations

The calculation of the power of a particular test proceeds as follows

1. Determine the critical value, $F_{df1,df2,\alpha}$ where *df1* is the numerator degrees of freedom, *df2* is the denominator degrees of freedom, and α is the probability of a type-I error (significance level). Note that the *F* test is a two-tailed test as no logical direction is assigned in the alternative hypothesis.

2. Calculate the standard deviation of the hypothesized effects, using the formula:

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}\left(e_i - \bar{e}\right)^2}{k}}$$

   where the e_i are effect values and *k* is the number of effects. Note that the average effect will be zero by construction, so this formula reduces to

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}\left(e_i\right)^2}{k}}$$

3. Compute the noncentrality parameter λ using the relationship:

$$\lambda = N \frac{\sigma_m^2}{\sigma^2}$$

   where N is the total number of subjects.

4. Compute the power as the probability of being greater than $F_{df1,df2,\alpha}$ on a noncentral-F distribution with noncentrality parameter λ.

Example

In the example discussed earlier, the standard deviation of the dose effects is

$$\sigma_m(A) = \sqrt{\frac{(-5.25)^2 + (-4.25)^2 + 9.50^2}{3}}$$
$$= 6.729908$$

the standard deviation of the diet effects is

$$\sigma_m(B) = \sqrt{\frac{(-3.5)^2 + 3.5^2}{2}}$$
$$= 3.5$$

and the standard deviation of the interaction effects is

$$\sigma_m(AB) = \sqrt{\frac{1.25^2 + (-1.25)^2 + 1.75^2 + (-1.75)^2 + (-3.00)^2 + 3.00^2}{6}}$$
$$= 2.131119$$

Change in Calculation from PASS 6.0

In ***PASS 6.0***, we used the approach of Cohen (1988) to calculate λ. However, we have found that Cohen's method is less accurate in some situations. Here's why. Cohen produced a set of tables for the one-way AOV which he extended to the two-way and three-way cases by adjusting the per group sample size (his n') so that the denominator degrees of freedom were accurate. However, his adjustment also causes a change in λ which can cause a substantial difference in the calculated power. By using the formula

$$\lambda = N \frac{\sigma_m^2}{\sigma^2}$$

we now calculate the correct power. This is why our calculations differ from that of Cohen (1988) for fixed factorial models.

Standard Deviation of Effects (of Means)

In the two-sample t-test case, the alternative hypothesis was represented as the difference between two group means. Unfortunately, for three or more groups, there is no simple extension of the two group difference. Instead, you must hypothesize a set of effects and calculate the value of σ_m.

Some might wish to specify the alternative hypothesis as the effect size, *f*, which is defined as

$$f = \frac{\sigma_m}{\sigma}$$

where σ is the standard deviation of values within a cell (see Sigma below). If you want to use *f*, set $\sigma = 1$ and then f is always equal to σ_m so that the values you enter for σ_m will be the values of *f*. Cohen (1988) has designated values of *f* less than 0.1 as *small*, values around 0.25 to be *medium*, and values over 0.4 to be *large*. You should come up with your own cutoff values for low, medium, and high.

When you are analyzing the power of an existing analysis of variance table, you can compute the values of σ_m for each term from its mean square or *F* ratio using the following formulas:

$$\sigma_m = \sqrt{\frac{df_{numerator} MS_{numerator}}{N}}$$

or

$$\sigma_m = \sqrt{\frac{df_{numerator}(F)(MSE)}{N}}$$

where *N* is the total number of observations, *MSE* is the mean square error, *df* is the numerator degrees of freedom, *MS* is the mean square of the term, and *F* is the *F* ratio of the term. If you do this, you are setting the sample effects equal to the population effects for the purpose of computing the power.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as the Template tab, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Sample Size

N per Cell

This is the sample size within a cell. Fractional values are allowed. When you have an unequal number of observations per cell, enter the average cell sample size.

If you enter more than one value, a separate analysis will be generated for each value.

Effect Size – Main Effects & Interactions

Factors (A, B, C) & Interactions (AB, …, ABC)

These check boxes specify which terms are included in the analysis of variance model. Check a term to signify that it must be included in the analysis.

The three factors are assigned the labels *A*, *B*, and *C*. The interaction between factors *A* and *B* is labeled *AB*. The three-way interaction is labeled *ABC*.

You cannot include an interaction term without including all shorter terms that make up that interaction. For example, if you include the interaction *AC*, you must also include the terms *A* and *C*. Similarly, if you include the term *ABC*, you must also include the terms *A, B, C, AB, AC*, and *BC*.

Effect Size – Main Effects

Categories (A, B, and C)

These options specify the number of categories (levels) contained in each factor. Since the total sample size is equal to the product of the number of levels in each factor and the number of observations per cell (N Per Cell), increasing the number of levels of a factor increases the total sample size of the experiment.

Hypothesized Means (A, B, C)

Enter a set of hypothesized means (or effects), one for each factor level. The standard deviation of these means is used in the power calculations. The standard deviation is calculated using the formula:

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{\left(\mu_i - \bar{\mu}\right)^2}{k}}$$

where *k* is the number of effects. Note that the standard deviation will be the same whether you enter means or effects since the average of the effects is zero by definition.

Enter a set of means that give the pattern of differences you expect or the pattern that you wish to detect. For example, in a particular study involving a factor with three categories, your research might be meaningful if either of two treatment means is 50% larger than the control mean. If the control mean is 50, then you would enter *50,75,75* as the three means.

It is usually more intuitive to enter a set of mean values. However, it is possible to enter the standard deviation of the means directly by placing an *S* in front of the number.

Entering a List of Means

If numbers are entered without a leading *S*, they are assumed to be the hypothesized group means under the alternative hypothesis. Their standard deviation will be calculated and used in the calculations. Blanks or commas may separate the numbers. Note that it is not the values of the means themselves that is important, but only their differences. Thus, the mean values *0,1,2* produce the same results as the values *100,101,102*.

If not enough means are entered to match the number of groups, the last mean is repeated. For example, suppose that four means are needed and you enter *1,2* (only two means). ***PASS*** will treat this as *1,2,2,2*. If too many values are entered, ***PASS*** will truncate the list to the number of means needed.

Examples:

5 20 60

2,5,7

-4,0,6,9

S Option

If an *S* is entered before a number, the number is assumed to be the value of σ_m, the standard deviation of the means.

Examples:

S 4.7

S 5.7

Effect Size – Interactions

Hypothesized Effects

Specify the standard deviation of the interaction effects using one of the following methods:

1. Enter a set of effects and let the program calculate their standard deviation (see below).
2. Enter the standard deviation directly.
3. Instruct the program to make the standard deviation proportional to one of the main effect terms.

The standard deviation of the effects is calculated using the formula:

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{\left(e_i - \bar{e}\right)^2}{k}}$$

$$= \sqrt{\sum_{i=1}^{k} \frac{e_i^2}{k}}$$

where k is the number of effects and $e_1, e_2, \cdots, e_k$ are the effect values. The value of $\bar{e}$ may be ignored because it is zero by definition.

Entering a List of Effects

If numbers are entered without a leading letter, they are assumed to be the hypothesized effects under the alternative hypothesis (they are all assumed to be zero under the null hypothesis). Their standard deviation will be calculated and used in the calculations. Blanks or commas may separate the numbers.

If not enough effects are entered to match the number of levels in the term, the last effect is repeated. For example, suppose that four effects are needed and you enter *1,2* (only two effects). ***PASS*** will treat this as *1,2,2,2*. If too many values are entered, ***PASS*** will truncate the list to the number of effects needed.

For interactions, the number of effects is equal to the product of the number of levels of each factor in the interaction. For example, suppose a two-factor design has one factor with three levels and another factor with five levels. The number of effects in the two-factor interaction is (3)(5) = 15.

Examples (note that they sum to zero):

-1 1 -3 3

2 2 0 -1 -1 -2

-4,0,1,3

S Option

If an *S* is followed by a number, the number is assumed to be the value of σ_m, the standard deviation of the effects.

When a set of effects are equal to either *e* or -*e*, the formula for the standard deviation may be simplified as follows:

$$\begin{aligned}\sigma_m &= \sqrt{\sum_{i=1}^{k}\frac{(e_i-0)^2}{k}}\\ &= \sqrt{\sum_{i=1}^{k}\frac{e^2}{k}}\\ &= e\end{aligned}$$

Hence, another interpretation of σ_m is the absolute value of a set of effects that are equal, except for the sign.

Example:

S 4.7

Enter a Term Followed by a Percentage

You can enter the name of a previous term followed by a percentage. This instructs the program to set this standard deviation to *x*% of the term you specify, where *x* is a positive integer. This allows you to set the magnitude of the interaction standard deviation as a percentage of another term without specifying the interaction in detail.

Note that the term you are taking a percentage of must appear above the term you are specifying. That is, you cannot specify *AB 50* for factor *C* (since only *A* and *B* occur above *C* on the screen).

For example, if the standard deviation of factor *A* is 16, the command

A 75

will set the standard deviation of the current term to (16)(75)/(100) = 12.0.

Other examples of this syntax are:

A 50

B 25

AB 125

AC 150

Discussion

The general formula for the calculation of the standard deviation is

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}\left(e_i\right)^2}{k}}$$

where *k* is the number of effects. In the case of a two-way interaction, the standard deviation is calculated using the formula:

$$\sigma_m\left(AB\right) = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \overline{\mu}\right)^2}{IJ}}$$

where *i* is the factor *A* index (from 1 to *I*), *j* is the factor *B* index (from 1 to *J*), μ_{ij} is the mean in the ij^{th} *cell*, $\mu_{i\bullet}$ is the i^{th} mean of factor *A* across all levels of other factors, $\mu_{\bullet j}$ is the j^{th} mean when factor *B* across all levels of other factors, and $\overline{\mu}$ is the overall mean of the means.

To see how this works, consider the following table of means from an experiment with *I* = 2 and *J* = 3:

		i		
		1	2	
	1	2.0	4.0	| 3.0
j	2	4.0	6.0	| 5.0
	3	6.0	11.0	| 8.5
	---	---	---	| ---
	Total	4.0	7.0	| 5.5

Now, if we subtract the factor *A* means, subtract the factor *B* means, and add the overall mean, we get the interaction effects:

0.5	-0.5
0.5	-0.5
-1.0	1.0

Next, we sum the squares of these six values:

$$\left(0.5\right)^2 + \left(-0.5\right)^2 + \left(0.5\right)^2 + \left(-0.5\right)^2 + \left(-1.0\right)^2 + \left(1.0\right)^2 = 3$$

Next we divide this value by (2)(3) = 6:

$$3/6 = 0.5$$

Finally, we take the square root of this value:

$$\sqrt{0.5} = 0.7071$$

Hence, for this configuration of means,

$$\sigma_m(AB) = 0.7071\,.$$

Notice that the average of the absolute values of the interaction effects is:

$$[0.5 + 0.5 + 0.5 + 0.5 + 1.0 + 1.0]/6 = 0.6667$$

We see that SD(interaction) is close to the average absolute interaction effect. That is, 0.7071 is close to 0.6667. This will usually be the case. Hence, one way to interpret the interaction standard deviation is as a number a little larger than the average absolute interaction effect.

Alpha

These options specify the significance levels (the probability of a type-I error) of each term. A type-I error occurs when you reject the null hypothesis of that all effects are zero when in fact they are.

Since they are probabilities, alpha values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This value may be interpreted as meaning that about one *F* test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You can select different alpha values for different terms. For example, although you have three factors in an experiment, you might be mainly interested in only one of them. Hence, you could increase the alpha level of the tests from, for example, 0.05 to 0.10 and thereby increase their power. Also, you may want to increase the alpha level of the interaction terms, since these will often have poor power otherwise.

Effect Size – Standard Deviation

S (Standard Deviation of Subjects)

This option specifies the value of the standard deviation (σ) within a cell (the analysis of variance assumes that σ is constant across all cells). Since they are positive square roots, the numbers must be strictly greater than zero. You can press the SD button to obtain further help on estimating the standard deviation.

This value may be estimated from a previous analysis of variance table by the square root of the mean square error.

If you want to use the effect size, *f*, as the measure of the variability of the effects, you can use 1.0 for σ.

Example 1 – Power after a Study

This example will explain how to calculate the power of *F* tests from data that have already been collected and analyzed.

Analyze the power of the experiment that was given at the beginning of this chapter. These data were analyzed using the analysis of variance procedure in ***NCSS*** and the following results were obtained.

```
Analysis of Variance Table
Source                      Sum of          Mean                          Prob          Power
Term              DF        Squares         Square          F-Ratio      Level         (Alpha=0.05)
A (Dose)           2        543.5           271.75          50.95        0.000172*     1.000000
B (Diet)           1        147             147             27.56        0.001920*     0.990499
AB                 2        54.5            27.25           5.11         0.050629      0.588884
S                  6        32              5.333333
Total (Adjusted)  11        777
Total             12
* Term significant at alpha = 0.05

Means and Effects Section
                                      Standard
Term          Count       Mean        Error          Effect
All              12       22.50                       22.50
A: Dose
High              4       32.00       1.154701         9.50
Medium            4       18.25       1.154701        -4.25
Low               4       17.25       1.154701        -5.25

B: Diet
D1                6       19.00       0.942809        -3.50
D2                6       26.00       0.942809         3.50

AB: Dose,Diet
High,D1           2       25.50       1.632993        -3.00
High,D2           2       38.50       1.632993         3.00
Low,D1            2       15.00       1.632993         1.25
Low,D2            2       19.50       1.632993        -1.25
Medium,D1         2       16.50       1.632993         1.75
Medium,D2         2       20.00       1.632993        -1.75
```

Setup

To analyze these data, we can enter the means for factors *A* and *B* as well as the *AB* interaction effects.

Alternatively, we could have calculated the standard deviation of the interaction. This can be done in either of two ways.

Using mean square for *AB* (27.25), the degrees of freedom for *AB* (2), and the total sample size (12), the standard deviation of the *AB*-interaction effects is calculated as follows

$$\sigma_m(AB) = \sqrt{\frac{2(27.25)}{12}} = 2.1311$$

Using the formula based on the effects, the standard deviation of the *AB*-interaction effects is calculated as follows

$$\sigma_m(AB) = \sqrt{\frac{3^2 + 3^2 + 1.25^2 + 1.25^2 + 1.75^2 + 1.75^2}{6}} = 2.1311$$

The value of σ is estimated from the square root of the mean square error:

$$\sigma = \sqrt{5.333333} = 2.3094$$

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Fixed Effects Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Fixed Effects ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
N Per Cell	**2**
Factors (A, B, AB)	***Checked***
Factors (C, AC, BC, ABC)	***Not checked***
Categories (A)	**3**
Categories (B)	**2**
Hypothesized Means (A)	**17.25 18.25 32**
Hypothesized Means (B)	**19 26**
Hypothesized Effects (AB)	**-3 3 1.25 -1.25 1.75 -1.75**
Alpha	***All are set to 0.05***
S (Std Dev of Subjects)	**2.3094**
Report Tab	
Report Prob Decimals	**6**
Std Dev Decimals	**4**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                      Std Dev
                              Total                  of Means    Effect
Term       Power        n         N     df1     df2      (Sm)      Size        Alpha         Beta
A       1.000000     2.00        12       2       6    6.7299     2.914     0.050000     0.000000
B       0.990499     2.00        12       1       6    3.5000     1.516     0.050000     0.009501
AB      0.588914     2.00        12       2       6    2.1311     0.923     0.050000     0.411086
Standard Deviation Within Subjects = 2.3094

Summary Statements
A factorial design with two factors at 3 and 2 levels has 6.0 cells (treatment combinations). A
total of 12.0 subjects are required to provide 2.0 subjects per cell. The within-cell standard
deviation is 2.3094. This design achieves 100% power when an F test is used to test factor A at
a 5% significance level and the actual standard deviation among the appropriate means is 6.7299
(an effect size of 2.914), achieves 99% power when an F test is used to test factor B at a 5%
significance level and the actual standard deviation among the appropriate means is 3.5000 (an
effect size of 1.516), and achieves 59% power when an F test is used to test the AB interaction
at a 5% significance level and the actual standard deviation among the appropriate means is
2.1312 (an effect size of 0.923).
```

This report shows the power for each of the three factors. Note that these power values match those given by the ***NCSS*** program in the analysis of variance report.

It is important to emphasize that these power values are for the case when the effects associated with the alternative hypotheses are equal to those given by the data. It will often be informative to calculate the power for other values as well.

Term

This is the term (main effect or interaction) from the analysis of variance model being displayed on this line.

Power

This is the power of the *F* test for this term. Note that since adding and removing terms changes the denominator degrees of freedom (*df2*), the power depends on which other terms are included in the model.

n

This is the sample size per cell (treatment combination). Fractional values indicate an unequal allocation among the cells.

Total N

This is the total sample size for the complete design.

df1

This is the numerator degrees of freedom of the *F* test.

df2

This is the denominator degrees of freedom of the *F* test. This value depends on which terms are included in the AOV model.

Std Dev of Means (Sm)

This is the standard deviation of the means (or effects). It represents the size of the differences among the effects that is to be detected by the analysis. If you have entered hypothesized means, only their standard deviation is displayed here.

Effect Size

This is the standard deviation of the means divided by the standard deviation of subjects. It provides an index of the magnitude of the difference among the means that can be detected by this design.

Alpha

This is the significance level of the *F* test. This is the probability of a type-I error given the null hypothesis of equal means and zero effects.

Beta

This is the probability of the type-II error for this test given the sample size, significance level, and effect size.

Example 2 – Finding the Sample Size

In this example, we will investigate the impact of increasing the sample size on the power of each of the seven tests in the analysis of variance table of a three factor experiment. The first factor (*A*) has two levels, the second factor (*B*) has three levels, and the third factor (*C*) has four levels. This creates a design with 2 x 3 x 4 = 24 treatment combinations.

All values of σ_m will be set equal to 0.2, σ is set equal to 1.0, and alpha is set to 0.05.

Setup

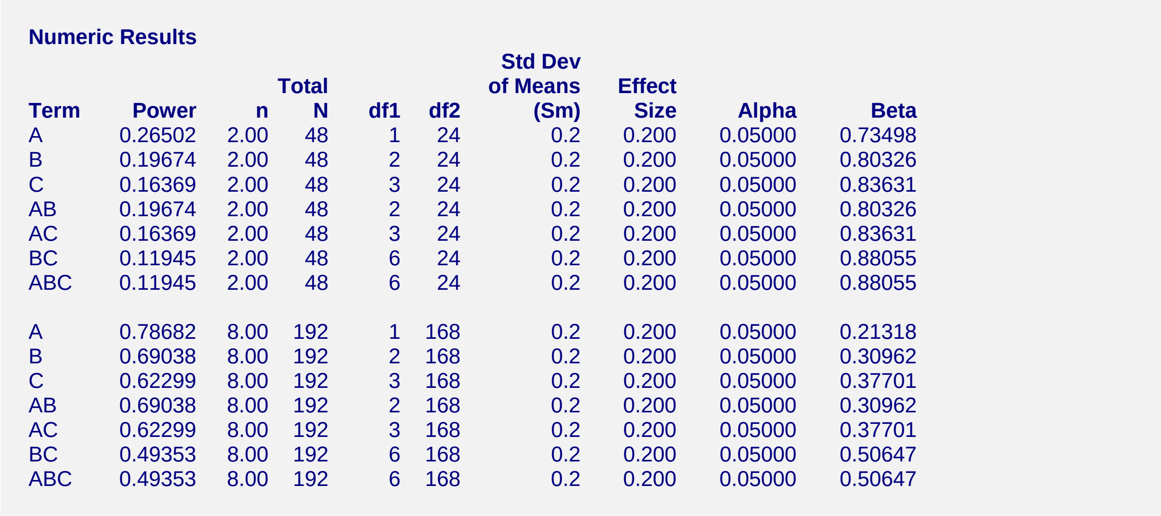
This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Fixed Effects Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Fixed Effects ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
N Per Cell	**2 8 16 22**
Factors and Interactions	***All Checked***
Categories (A)	**2**
Categories (B)	**3**
Categories (C)	**4**
Hypothesized Means (A, B, & C)	**S 0.2**
Hypothesized Effects (AB to ABC)	**A 100 *(so they will equal that of factor A)***
Alpha	***All are set to 0.05***
S (Std Dev of Subjects)	**1.0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results



```
Numeric Results
                                                 Std Dev
                          Total                 of Means   Effect
Term      Power      n       N   df1   df2          (Sm)     Size     Alpha       Beta
A       0.26502   2.00      48     1    24           0.2    0.200   0.05000    0.73498
B       0.19674   2.00      48     2    24           0.2    0.200   0.05000    0.80326
C       0.16369   2.00      48     3    24           0.2    0.200   0.05000    0.83631
AB      0.19674   2.00      48     2    24           0.2    0.200   0.05000    0.80326
AC      0.16369   2.00      48     3    24           0.2    0.200   0.05000    0.83631
BC      0.11945   2.00      48     6    24           0.2    0.200   0.05000    0.88055
ABC     0.11945   2.00      48     6    24           0.2    0.200   0.05000    0.88055

A       0.78682   8.00     192     1   168           0.2    0.200   0.05000    0.21318
B       0.69038   8.00     192     2   168           0.2    0.200   0.05000    0.30962
C       0.62299   8.00     192     3   168           0.2    0.200   0.05000    0.37701
AB      0.69038   8.00     192     2   168           0.2    0.200   0.05000    0.30962
AC      0.62299   8.00     192     3   168           0.2    0.200   0.05000    0.37701
BC      0.49353   8.00     192     6   168           0.2    0.200   0.05000    0.50647
ABC     0.49353   8.00     192     6   168           0.2    0.200   0.05000    0.50647
```

```
A      0.97434  16.00  384    1  360      0.2   0.200   0.05000   0.02566
B      0.94723  16.00  384    2  360      0.2   0.200   0.05000   0.05277
C      0.92061  16.00  384    3  360      0.2   0.200   0.05000   0.07939
AB     0.94723  16.00  384    2  360      0.2   0.200   0.05000   0.05277
AC     0.92061  16.00  384    3  360      0.2   0.200   0.05000   0.07939
BC     0.84559  16.00  384    6  360      0.2   0.200   0.05000   0.15441
ABC    0.84559  16.00  384    6  360      0.2   0.200   0.05000   0.15441

A      0.99569  22.00  528    1  504      0.2   0.200   0.05000   0.00431
B      0.98880  22.00  528    2  504      0.2   0.200   0.05000   0.01120
C      0.98045  22.00  528    3  504      0.2   0.200   0.05000   0.01955
AB     0.98880  22.00  528    2  504      0.2   0.200   0.05000   0.01120
AC     0.98045  22.00  528    3  504      0.2   0.200   0.05000   0.01955
BC     0.95001  22.00  528    6  504      0.2   0.200   0.05000   0.04999
ABC    0.95001  22.00  528    6  504      0.2   0.200   0.05000   0.04999

Standard Deviation of Subjects = 1.0
```

A few interesting features of this report stand out. First note the range of power values across the range of sample size values tested. Reasonable power is not reached until *n* is 16. Also note that as the number of numerator degrees of freedom (*df1*) increases, the power decreases, other things being equal. We must use this knowledge when planning for appropriate power in tests of important interaction terms.

There are a lot of additional runs that you might try. For example, you might look at the impact of setting the alpha level of interaction terms 0.08. You might look at varying σ_m across the different terms. You might try varying the number of levels of a factor. All of these will impact the power of the *F* tests and will thus be important to consider during the planning stage of an experiment.

Example 3 – Latin Square Design

This example shows how to study the power of a complicated experimental design like a Latin square. Suppose you want to run a Five-Level Latin square design. Recall that a Five-Level Latin square design consists of three factors each at five levels. One factor is associated with the columns of the square, a second factor is associated with the rows of the square, and a third factor is associated with the letters of the square. In all there are only 5 x 5 = 25 observations used instead of the 5 x 5 x 5 = 125 that would normally be required. The Latin square design has reduced the number of observations by 80%.

The 80% decrease in observations comes at a price—the interaction terms must be ignored. If you can legitimately assume that the interactions are zero, the Latin square (or some other design which reduces the number of observations) is an efficient design to use. We will now show you how to analyze the power of the *F* tests from such a design.

The key is to enter 0.2 (which is 25/125) for *n* and set all the interaction indicators off.

Since all three factors have five levels, the power of the three *F* tests will be the same if σ_m is the same. Hence, we can try three different sets of hypothesized means. The first set will be five means 0.1 units apart. The second set will be five means 0.5 units apart. The third set will be five means 1.0 unit apart. The standard deviation will be set to 1.0. All alpha levels will be set at 0.05.

The sample size per cell is set at 0.2 and 0.4. This will result in total sample sizes of 25 (one replication) and 50 (two replications).

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Fixed Effects Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Fixed Effects ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
N Per Cell	**0.2 0.4**
Factors (A, B, C)	***Checked***
Interactions (AB, AC, BC, ABC)	***Not checked***
Categories (A, B, C)	**5**
Hypothesized Means (A)	**1.0 1.1 1.2 1.3 1.4**
Hypothesized Means (B)	**1.0 1.5 2.0 2.5 3.0**
Hypothesized Means (C)	**1.0 2.0 3.0 4.0 5.0**
Alpha	***All are set to 0.05***
S (Std Dev of Subjects)	**1.0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                  Std Dev
                         Total                   of Means    Effect
Term      Power       n      N    df1    df2        (Sm)      Size      Alpha       Beta
A       0.06807    0.20     25      4     12       0.141     0.141    0.05000    0.93193
B       0.63675    0.20     25      4     12       0.707     0.707    0.05000    0.36325
C       0.99867    0.20     25      4     12       1.414     1.414    0.05000    0.00133

A       0.09842    0.40     50      4     37       0.141     0.141    0.05000    0.90158
B       0.97743    0.40     50      4     37       0.707     0.707    0.05000    0.02257
C       1.00000    0.40     50      4     37       1.414     1.414    0.05000    0.00000

Standard Deviation of Subjects = 1.000
```

In the first design in which *N* = 25, only the power of the test for *C* is greater than 0.8. Of course, this power value also depends on the value of the standard deviation of subjects within a cell.

It is interesting to note that doubling the sample size did not double the power!

Example 4 – Validation using Winer

Winer (1991) pages 428-429 presents the power calculations for a two-way design in which factor *A* has two levels and factor *B* has three levels. Winer provides estimates of the sum of squared *A* effects (1.0189), sum of squared *B* effects (5.06), and sum of squared interaction effects (42.11). The mean square error is 8.83 and the per cell sample size is 3. All alpha levels are set to 0.05.

Winer's results are approximate because he has to interpolate in the tables that he is using. He finds the power of the *F* test for factor *A* to be between 0.10 and 0.26. He estimates it as 0.17. The exact power of the *F* test for factor *B* is not given. Instead, the range is found to be between 0.26 and 0.36. The power of the *F* test for the *AB* interaction is "approximately" 0.86.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Fixed Effects Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Fixed Effects ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
N Per Cell	**3**
Factors (A and B)	***Checked***
Factor (C)	***Not checked***
Interaction (AB)	***Checked***
Interactions (AC, BC, ABC)	***Not checked***
Categories (A)	**2**
Categories (B)	**3**
Hypothesized Means (A)	**S 0.714**
Hypothesized Means (B)	**S 1.3**
Hypothesized Effects (AB)	**S 2.65**
Alpha	***All are set to 0.05***
S (Std Dev of Subjects)	**2.97**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                    Std Dev
                             Total                 of Means   Effect
Term      Power        n        N    df1    df2       (Sm)     Size      Alpha        Beta
A         0.15576   3.00       18      1     12      0.714    0.240    0.05000     0.84424
B         0.29178   3.00       18      2     12      1.300    0.438    0.05000     0.70822
AB        0.85338   3.00       18      2     12      2.650    0.892    0.05000     0.14662

Standard Deviation of Subjects = 2.970
```

The power of the test for factor A is 0.16 which is between 0.10 and 0.26. It is close to the interpolated 0.17 that Winer obtained from his tables.

The power of the test for factor B is 0.29 which is between 0.26 and 0.36.

The power of the test for the AB interaction is 0.85 which is close to the interpolated 0.86 that Winer obtained from his tables.

Example 5 – Validation using Prihoda

Prihoda (1983) pages 7-8 presents the power calculations for a two-way design with the following pattern of means:

		Factor B				
		1	2	3	4	All
Factor A	1	41	34	30	27	33
	2	33	24	22	29	27
All		37	29	26	28	30

The means may be manipulated to show the overall mean, the main effects, and the interaction effects:

		Factor B				
		1	2	3	4	All
Factor A	1	1	2	1	-4	3
	2	-1	-2	-1	4	-3
All		7	-1	-4	-2	30

Based on the above effects, Prihoda calculates the power of the interaction test when the sample size per cell is 6, 8, 10, 12, and 14 to be 0.34, 0.45, 0.56, 0.65, and 0.73. The mean square error is 64 and the alpha level is 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Fixed Effects Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Fixed Effects ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
N Per Cell	**6 8 10 12 14**
Factors (A and B)	***Checked***
Factor (C)	***Not checked***
Interaction (AB)	***Checked***
Interactions (AC, BC, ABC)	***Not checked***
Categories (A)	**2**
Categories (B)	**4**
Hypothesized Means (A)	**33 27**
Hypothesized Means (B)	**37 29 26 28**
Hypothesized Effects (AB)	**1 -2 2 -2 1 -1 -4 4**
Alpha	***All are set to 0.05***
S (Std Dev of Subjects)	**8**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                    Std Dev
                             Total                 of Means   Effect
Term     Power        n          N    df1    df2       (Sm)     Size       Alpha        Beta
A       0.71746    6.00         48      1     40      3.000    0.375     0.05000     0.28254
B       0.83676    6.00         48      3     40      4.183    0.523     0.05000     0.16324
AB      0.33722    6.00         48      3     40      2.345    0.293     0.05000     0.66278

A       0.83848    8.00         64      1     56      3.000    0.375     0.05000     0.16152
B       0.93871    8.00         64      3     56      4.183    0.523     0.05000     0.06129
AB      0.45099    8.00         64      3     56      2.345    0.293     0.05000     0.54901

A       0.91134   10.00         80      1     72      3.000    0.375     0.05000     0.08866
B       0.97917   10.00         80      3     72      4.183    0.523     0.05000     0.02083
AB      0.55558   10.00         80      3     72      2.345    0.293     0.05000     0.44442

A       0.95292   12.00         96      1     88      3.000    0.375     0.05000     0.04708
B       0.99346   12.00         96      3     88      4.183    0.523     0.05000     0.00654
AB      0.64749   12.00         96      3     88      2.345    0.293     0.05000     0.35251

A       0.97568   14.00        112      1    104      3.000    0.375     0.05000     0.02432
B       0.99807   14.00        112      3    104      4.183    0.523     0.05000     0.00193
AB      0.72541   14.00        112      3    104      2.345    0.293     0.05000     0.27459

Standard Deviation of Subjects = 8.000
```

Prihoda only presents the power for the interaction test at each sample size. You can check to see that the results match Prihoda's exactly.

Example 6 – Validation using Neter, Kutner, Nachtsheim, and Wasserman

Neter, Kutner, Nachtsheim, and Wasserman (1996) page 1057 presents a power analysis of a two-factor experiment in which factor *A* has three levels and factor *B* has two levels. The significance level is 0.05, the standard deviation is 3.0, and *N* is 2. They calculate a power of about 0.89 for the test of factor *A* when the three means are 50, 55, and 45.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Fixed Effects Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Fixed Effects ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
N Per Cell	**2**
Factors (A and B)	***Checked***
Factor (C)	***Not checked***
Interaction (AB)	***Checked***
Interactions (AC, BC, ABC)	***Not checked***
Categories (A)	**3**
Categories (B)	**2**
Hypothesized Means (A)	**50 55 45**
Hypothesized Means (B)	**S 1**
Hypothesized Effects (AB)	**S 1**
Alpha	***All are set to 0.05***
S (Std Dev of Subjects)	**3.0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                 Std Dev
                                 Total          of Means   Effect
Term        Power        n          N   df1  df2    (Sm)     Size      Alpha       Beta
A         0.90162     2.00         12     2    6   4.082    1.361    0.05000    0.09838
B         0.16479     2.00         12     1    6   1.000    0.333    0.05000    0.83521
AB        0.11783     2.00         12     2    6   1.000    0.333    0.05000    0.88217

Standard Deviation of Subjects = 3.000
```

Note that the power of 0.90 that ***PASS*** has calculated is within rounding of the 0.89 that Neter *et al.* calculated.

Chapter 565

Randomized Block Analysis of Variance

Introduction

This module analyzes a randomized block analysis of variance with up to two treatment factors and their interaction. It provides tables of power values for various configurations of the randomized block design.

The Randomized Block Design

The randomized block design (*RBD*) may be used when a researcher wants to reduce the experimental error among observations of the same treatment by accounting for the differences among blocks. If three treatments are arranged in two blocks, the *RBD* might appear as follows:

Block A	**Block B**
Treatment 1	Treatment 2
Treatment 3	Treatment 1
Treatment 2	Treatment 3

This diagram shows the main features of a *RBD*:

1. Each block is divided into *k* sub-blocks, where *k* is the number of treatments.
2. Each block receives all the treatments.
3. The treatments are assigned to the sub-blocks in random order.
4. There is some reason to believe that the blocks are the same internally, but different from each other.

RBD Reduces Random Error

The random error component of a completely randomized design (such as a one-way or a fixed-effects factorial design) represents the influence of all possible variables in the universe on the response except for the controlled (treatment) variables. This random error component is called the standard deviation or σ (sigma).

As we have discussed, the sample size required to meet alpha and beta error requirements depends directly on the standard deviation. As the standard deviation increases, the sample size increases. Hence, researchers are always looking for ways to reduce the standard deviation. Since the random error component contains the variation due to all possible variables other than treatment variables, one of the most obvious ways to reduce the standard deviation is to remove one or more of these *nuisance* variables from the random error component. One of the simplest ways of doing this is by blocking on them.

For example, an agricultural experiment is often blocked on fields so that differences among fields are explicitly accounted for and removed from the error component. Since these field differences are caused by variations in variables such as soil type, sunlight, temperature, and water, blocking on fields removes the influence of several variables.

Blocks are constructed so that the response is as alike (homogeneous) as possible within a block, but as different as possible between blocks. In many situations, there are obvious natural blocking factors such as schools, seasons, individual farms, families, times of day, etc. In other situations, the blocks may be somewhat artificially constructed.

Once the blocks are defined, they are divided into *k* smaller sections called *subblocks*, where *k* is the number of treatment levels. The *k* treatments are randomly assigned to the subblocks, one block at a time. Hence the order of treatment application will be different from block to block.

Measurement of Random Error

The measurement of the random error component (σ) is based on the assumption that there is no fundamental relationship between the treatment variable and the blocking variable. When this is true, the interaction component between blocks and treatment is zero. If the interaction component is zero, then the amount measured by the interaction is actually random error and can be used as an estimate of σ.

Hence, the randomized block design makes the assumption that there is no interaction between treatments and blocks. The block by treatment mean square is still calculated, but it is used as the estimated standard deviation. This means that the degrees of freedom associated with the block-treatment interaction are the degrees of freedom of the error estimate. If the experimental design has *k* treatments and *b* blocks, the interaction degrees of freedom are equal to *(k-1)(b-1)*. Hence the sample size of this type of experiment is measured in terms of the number of blocks.

Treatment Effects

Either one or two treatment variables may be specified. If two are used, their interaction may also be measured. The null hypothesis in the *F* test states that the effects of the treatment variable are zero. The magnitude of the alternative hypothesis is represented as the size of the standard deviation (σ_m) of these effects. The larger the size of the effects, the larger their standard deviation.

When there are two factors, the block-treatment interaction may be partitioned just as the treatment may be partitioned. For example, if we let *C* and *D* represent two treatments, an analysis of variance will include the terms *C*, *D*, and *CD*. If we represent the blocking factor as *B*, there will be three interactions with blocks: *BC*, *BD*, and *BCD*. Since all three of these terms are assumed to measure the random error, the overall estimate of random error is found by averaging (or *pooling*) these three interactions. The pooling of these interactions increases the power of the experiment by effectively increasing the sample size on which the estimate of σ is based. However, it is based on the assumption that $\sigma = \sigma_{BC} = \sigma_{BD} = \sigma_{BCD}$, which may or may not be true.

An Example

Following is an example of data from a randomized block design. The block factor has four blocks (*B1*, *B2*, *B3*, *B4*) while the treatment factor has three levels (low, medium, and high). The response is shown within the table.

Randomized Block Example

	Treatments		
Blocks	**Low**	**Medium**	**High**
B1	16	19	20
B2	18	20	21
B3	15	17	22
B4	14	17	19

Analysis of Variance Hypotheses

The *F* test for treatments in a randomized block design tests the hypothesis that the treatment effects are zero. (See the beginning of the Fixed-Effects Analysis of Variance chapter for a discussion of the meaning of effects.)

Single-Factor Repeated Measures Designs

The randomized block design is often confused with a single-factor repeated measures design because the analysis of each is similar. However, the randomization pattern is different. In a randomized block design, the treatments are applied in random order within each block. In a repeated measures design, however, the treatments are usually applied in the same order through time. You should not mix the two. If you are analyzing a repeated measures design, we suggest that you use that module of ***PASS*** to do the sample size and power calculations.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as the Template tab, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Sample Size

Number of Blocks

This specifies one or more values for the number of blocks. If a list of values is entered, a separate calculation will be made for each value.

Effect Size – Main Effects & Interactions

Factors (A, B, and AB)

These check boxes specify which terms are included in the analysis of variance model. Check a term to indicate that it is included.

The two factors are assigned the labels *A* and *B*. The interaction between factors *A* and *B* is labeled *AB*. You cannot include the interaction term without including both *A* and *B*.

Effect Size – Main Effects

Categories (A and B)

This option specifies the number of categories (levels) contained in each factor. Since the effective sample size is equal to the product of the number of levels in each factor and the number of blocks, increasing the number of levels of a factor increases the sample size of the experiment.

Hypothesized Means (A and B)

Enter a set of hypothesized means (or effects), one for each factor level. The standard deviation of these means is used in the power calculations. The standard deviation is calculated using the formula:

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{(e_i - \bar{e})^2}{k}}$$

where *k* is the number of levels. Note that the standard deviation will be the same whether you enter means or effects since the average of the effects is zero by definition.

Enter a set of means that give the pattern of differences you expect or the pattern that you wish to detect. For example, in a particular study involving a factor with three categories, your research might be meaningful if either of two treatment means is 50% larger than the control mean. If the control mean is 50, then you would enter *50,75,75* as the three means.

It is usually more intuitive to enter a set of mean values. However, it is possible to enter the standard deviation of the means directly by placing an *S* in front of the number.

Entering a List of Means

If numbers are entered without a leading *S*, they are assumed to be the hypothesized group means under the alternative hypothesis. Their standard deviation will be calculated and used in the calculations. Blanks or commas may separate the numbers. Note that it is not the values of the means themselves that is important, but only their differences. Thus, the mean values *0,1,2* produce the same results as the values *100,101,102*.

If not enough means are entered to match the number of groups, the last mean is repeated. For example, suppose that four means are needed and you enter *1,2* (only two means). ***PASS*** will treat this as *1,2,2,2*. If too many values are entered, ***PASS*** will truncate the list to the number of means needed.

Examples:

5 20 60

2,5,7

-4,0,6,9

S Option

If an *S* is entered before a number, the number is assumed to be the value of σ_m, the standard deviation of the means.

Examples:

S 4.6

S 5.8

Effect Size – Interactions

Hypothesized Effects

Specify the standard deviation of the interaction effects using one of the following methods:

1. Enter a set of effects and let the program calculate their standard deviation.
2. Enter the standard deviation directly.
3. Instruct the program to make the standard deviation proportional to one of the main effect terms.

The standard deviation of the effects is calculated using the formula:

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{(e_i - \bar{e})^2}{k}}$$

where k is the number of effects and $e_1, e_2, \cdots, e_k$ are the effect values. The value of $\bar{e}$ may be ignored because it is zero by definition.

Entering a List of Effects

If numbers are entered without a leading letter, they are assumed to be the hypothesized effects under the alternative hypothesis (they are all assumed to be zero under the null hypothesis). Their standard deviation will be calculated and used in the calculations. Blanks or commas may separate the numbers.

If not enough effects are entered to match the number of levels in the term, the last effect is repeated. For example, suppose that four effects are needed and you enter *1,2* (only two effects). ***PASS*** will treat this as *1,2,2,2*. If too many values are entered, ***PASS*** will truncate the list to the number of effects needed.

For interactions, the number of effects is equal to the product of the number of levels of each factor in the interaction. For example, suppose a two-factor design has one factor with three levels and another factor with five levels. The number of effects in the two-factor interaction is (3)(5) = 15.

Examples (note that they sum to zero):

-1 1 -3 3

2 2 0 -1 -1 -2

-4,0,1,3

S Option

If an *S* is followed by a number, the number is assumed to be the value of σ_m, the standard deviation of the effects.

When a set of effects are equal to either *e* or -*e*, the formula for the standard deviation may be simplified as follows:

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{(e_i - \bar{e})^2}{k}}$$

Hence, another interpretation of σ_m is the absolute value of a set of effects that are equal, except for the sign.

Example:

S 4.7

Enter a Term Followed by a Percentage

You can enter the name of a previous term followed by a percentage. This instructs the program to set this standard deviation to *x*% of the term you specify, where *x* is a positive integer. This allows you to set the magnitude of the interaction standard deviation as a percentage of the standard deviation of one of the factors without specifying the interaction in detail.

For example, if the standard deviation of factor *A* is 16, the command

A 75

will set the standard deviation of the current term to (16)(75)/(100) = 12.0.

Other examples of this syntax are:

A 50

B 25

Discussion

The general formula for the calculation of the standard deviation is

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k} (e_i)^2}{k}}$$

where k is the number of effects. In the case of a two-way interaction, the standard deviation is calculated using the formula:

$$\sigma_m(AB) = \sqrt{\frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \overline{\mu}\right)^2}{IJ}}$$

where i is the factor A index (from 1 to I), j is the factor B index (from 1 to J), μ_{ij} is the mean in the ij^{th} *cell,* $\mu_{i\bullet}$ is the i^{th} mean of factor A across all levels of other factors, $\mu_{\bullet j}$ is the j^{th} mean when factor B across all levels of other factors, and $\overline{\mu}$ is the overall mean of the means.

To see how this works, consider the following table of means from an experiment with I = 2 and J = 3:

		i		
		1	2	
	1	2.0	4.0	| 3.0
j	2	4.0	6.0	| 5.0
	3	6.0	11.0	| 8.5
	---	---	---	| ---
	Total	4.0	7.0	| 5.5

Now, if we subtract the factor A means, subtract the factor B means, and add the overall mean, we get the interaction effects:

0.5	-0.5
0.5	-0.5
-1.0	1.0

Next, we sum the squares of these six values:

$$(0.5)^2 + (-0.5)^2 + (0.5)^2 + (-0.5)^2 + (-1.0)^2 + (1.0)^2 = 3$$

Next we divide this value by (2)(3) = 6:

$$3/6 = 0.5$$

Finally, we take the square root of this value:

$$\sqrt{0.5} = 0.7071$$

Hence, for this configuration of means,

$$\sigma_m(AB) = 0.7071.$$

Notice that the average of the absolute values of the interaction effects is:

$$[0.5 + 0.5 + 0.5 + 0.5 + 1.0 + 1.0]/6 = 0.6667.$$

We see that SD(interaction) is close to the average absolute interaction effect. That is, 0.7071 is close to 0.6667. This will usually be the case. Hence, one way to interpret the interaction standard deviation is as a number a little larger than the average absolute interaction effect.

Alpha

This option specifies the probability of a type-I error (alpha) for each term. A type-I error occurs when you reject the null hypothesis that the effects are zero when in fact they are.

Since they are probabilities, alpha values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This value may be interpreted as meaning that about one *F* test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You can select different alpha values for different terms. For example, although you have three factors in an experiment, you might be mainly interested in only one of them. Hence, you could increase the alpha level of the tests from, for example, 0.05 to 0.10 and thereby increase their power. Also, you may want to increase the alpha level of the interaction terms, since these will often have poor power otherwise.

S (Standard Deviation)

This option specifies the value of the standard deviation. In a randomized block design, this value is estimated by the square root of the mean square error (which may be listed as the mean square of the block-by-treatment interaction). This value will usually have to be determined from a previous study.

Assuming that each block is divided into several subblocks, this is an estimate of the standard deviation that would result when the subblocks within the same block received the same treatment.

If you want to use the effect size, *f*, as the measure of the variability of the effects, you can use 1.0 for σ.

Estimation of the standard deviation is discussed in detail in the Standard Deviation Estimator chapter.

Example 1 – Power after a Study

This example will explain how to calculate the power of *F* tests from data that have already been collected and analyzed.

We will analyze the power of the experiment that was given at the beginning of this chapter.

These data were analyzed using the analysis of variance procedure in ***NCSS*** and the following results were obtained.

```
Analysis of Variance Table
Source                        Sum of        Mean                      Prob          Power
Term                DF        Squares       Square       F-Ratio      Level         (Alpha=0.05)
A (Blocks)           3        13.66667      4.555555
B (Treatment)        2        45.16667      22.58333     19.83        0.002269*     0.991442
AB                   6        6.833333      1.138889
S                    0        0
Total (Adjusted)    11        65.66666
Total               12
* Term significant at alpha = 0.05

Means and Effects Section
                                  Standard
Term        Count      Mean      Error                   Effect
B: Treatment
High            4      20.5      0.5335937             2.333333
Low             4     15.75      0.5335937            -2.416667
Medium          4     18.25      0.5335937        8.333334E-02
```

We will now calculate the power of the *F* test. Note that factor *B* in this printout becomes factor *A* on the ***PASS*** template.

Setup

To analyze these data, we enter the means for factor *A*. The value of σ is estimated as the square root of the mean square error:

$$\sigma = \sqrt{1.138889} = 1.0672$$

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Randomized Block Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Randomized Block ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Number of Blocks	**2 3 4 5**
Factor (A)	***Checked***
Factors (B, AB)	***Not checked***
Categories (A)	**3**
Hypothesized Means	**15.75 18.25 20.50**
Alpha	***All are set to 0.05***
S (Standard Deviation)	**1.0672**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                        Std Dev
                                                       of Means    Effect
Term    Power     Blocks   Units    df1     df2       (Sm)       Size       Alpha        Beta
A       0.42132        2       6      2       2      1.940     1.8179     0.05000     0.57868
A       0.89376        3       9      2       4      1.940     1.8179     0.05000     0.10624
A       0.99144        4      12      2       6      1.940     1.8179     0.05000     0.00856
A       0.99956        5      15      2       8      1.940     1.8179     0.05000     0.00044
Standard Deviation Within Blocks (block-treatment interaction) = 1.067


Summary Statements
A randomized-block design with one treatment factor at 3 levels has 2.0 blocks each with 3.0
treatment combinations. The square root of the block-treatment interaction is 1.067. This
design achieves 42% power when an F test is used to test factor A at a 5% significance level
and the actual standard deviation among the appropriate means is 1.940 (an effect size of
1.8179).
```

This report shows the power for each of the five block counts. We see that adequate power of about 0.9 would have been achieved by three blocks.

It is important to emphasize that these power values are for the case when the effects associated with the alternative hypotheses are equal to those given by the data. It will often be informative to calculate the power for other values as well.

Term

This is the term (main effect or interaction) from the analysis of variance model being displayed on this line.

Power

This is the power of the *F* test for this term. Note that since adding and removing terms changes the denominator degrees of freedom (*df2*), the power depends on which other terms are included in the model.

Blocks

This is the number of blocks in the design.

Units

This is the number of subblocks (plots) in the design. It is the product of the number of treatment levels and the number of blocks.

df1

This is the numerator degrees of freedom of the *F* test.

df2

This is the denominator degrees of freedom of the *F* test. This value depends on which terms are included in the AOV model.

Std Dev of Means (Sm)

This is the standard deviation of the means (or effects). It represents the size of the differences among the effects that is to be detected by the analysis. If you have entered hypothesized means, only their standard deviation is displayed here.

Effect Size

This is the standard deviation of the means divided by the standard deviation of subjects. It provides an index of the magnitude of the difference among the means that can be detected by this design.

Alpha

This is the significance level of the *F* test. This is the probability of a type-I error given the null hypothesis of equal means and zero effects.

Beta

This is the probability of the type-II error for this test given the sample size, significance level, and effect size.

Example 2 – Validation using Prihoda

Prihoda (1983) presents details of an example that is given in Odeh and Fox (1991). In this example, *Alpha* is 0.025, *Sm* of *A* is 0.577, the number of treatments in factor *A* is 6, the number of treatments in factor *B* is 3, S is 1.0, and the Number of Blocks is 2, 3, 4, 5, 6, 7, and 8. Prihoda gives the power values for the *F* test on factor *A* as 0.477, 0.797, 0.935, 0.982, 0.995, 0.999, and 1.000.

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Randomized Block Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Randomized Block ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Number of Blocks	**2 3 4 5 6 7 8**
Terms (A, B, AB)	***Checked***
Categories (A)	**6**
Categories (B)	**3**
Hypothesized Means (A)	**S 0.577**
Hypothesized Means (B)	**S 1**
Hypothesized Effects (AB)	**S 1**
Alpha	***All are set to 0.025***
S (Standard Deviation)	**1.0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results
                                                    Std Dev
                                                   of Means    Effect
Term    Power    Blocks   Units    df1     df2       (Sm)       Size      Alpha       Beta
A      0.47622        2      36      5      17      0.577     0.5770    0.02500    0.52378
B      0.99697        2      36      2      17      1.000     1.0000    0.02500    0.00303
AB     0.85337        2      36     10      17      1.000     1.0000    0.02500    0.14663

A      0.79521        3      54      5      34      0.577     0.5770    0.02500    0.20479
B      0.99999        3      54      2      34      1.000     1.0000    0.02500    0.00001
AB     0.99615        3      54     10      34      1.000     1.0000    0.02500    0.00385

A      0.93479        4      72      5      51      0.577     0.5770    0.02500    0.06521
B      1.00000        4      72      2      51      1.000     1.0000    0.02500    0.00000
AB     0.99995        4      72     10      51      1.000     1.0000    0.02500    0.00005

A      0.98226        5      90      5      68      0.577     0.5770    0.02500    0.01774
B      1.00000        5      90      2      68      1.000     1.0000    0.02500    0.00000
AB     1.00000        5      90     10      68      1.000     1.0000    0.02500    0.00000

A      0.99573        6     108      5      85      0.577     0.5770    0.02500    0.00427
B      1.00000        6     108      2      85      1.000     1.0000    0.02500    0.00000
AB     1.00000        6     108     10      85      1.000     1.0000    0.02500    0.00000

A      0.99907        7     126      5     102      0.577     0.5770    0.02500    0.00093
B      1.00000        7     126      2     102      1.000     1.0000    0.02500    0.00000
AB     1.00000        7     126     10     102      1.000     1.0000    0.02500    0.00000

A      0.99981        8     144      5     119      0.577     0.5770    0.02500    0.00019
B      1.00000        8     144      2     119      1.000     1.0000    0.02500    0.00000
AB     1.00000        8     144     10     119      1.000     1.0000    0.02500    0.00000

Standard Deviation Within Blocks (block-treatment interaction) = 1.000
```

We have bolded the power values on this report that should match Prihoda’s results. You see that they do match.

Chapter 570

Repeated Measures Analysis of Variance

Introduction

This module calculates the power for *repeated measures* designs having up to three within factors and up to three between factors. It computes power for various test statistics including the *F* test with the Geisser-Greenhouse correction, Wilks' lambda, Pillai-Bartlett trace, and Hotelling-Lawley trace. It can be used to calculate the power of *crossover* designs. .

Repeated measures designs are popular because they allow a subject to serve as their own control. This usually improves the precision of the experiment. However, when the analysis of the data uses the traditional *F* tests, additional assumptions concerning the structure of the error variance must be made. When these assumptions do not hold, the Geisser-Greenhouse correction provides reasonable adjustments so that significance levels are accurate.

An alternative to using the *F* test with repeated measures designs is to use one of the multivariate tests: Wilks' lambda, Pillai-Bartlett trace, or Hotelling-Lawley trace. These alternatives are appealing because they do not make the strict, often unrealistic, assumptions about the structure of the error variance. Unfortunately, they may have less power than the *F* test and they cannot be used in all situations.

An example of a two-factor repeated measures design that can be analyzed by this procedure is shown by the following diagram.

Group 1			**Group 2**	
Subject 1	**Subject 2**	**Month**	**Subject 3**	**Subject 4**
Treatment L	Treatment L	1	Treatment L	Treatment L
Treatment M	Treatment M	2	Treatment M	Treatment M
Treatment H	Treatment H	3	Treatment H	Treatment H

Groups 1 and 2 form the *between* factor. The within factor has three levels: *L*, *M*, and *H* (low, medium, and high). There are four subjects in this experiment. The three treatments are applied to each subject, one treatment per month. Note that the three treatments are applied to each subject in the same order. Although the order of treatment application should be randomized, it is often the same for all subjects.

This diagram shows the main features of a repeated measures design, which are

1. Each subject receives all treatments.
2. The treatments are applied through time. When the treatments are applied in the same order across all subjects, it is impossible to separate the treatment effects from the sequence effects. Some processes that can cause *sequence effects* are learning, practice, or fatigue—any pattern in the responses across time that occurs without the treatment. If you think the possibility for sequence effects exists, you must make sure that the effects of prior treatments have been washed out before applying the next treatment.
3. Unlike other designs, the repeated measures design has two experimental units: *between* and *within*. In this example, the first (between*)* experimental unit is a subject. Subject-to-subject variability is used to test the between factor (groups). The second (within) experimental unit is the time period. In the above example, the month to month variability within a subject is used to test the treatment. The important point to realize is that the repeated measures design has two error components, the between and the within.

Assumptions

The following assumptions are made when using the *F* test to analyze a factorial experimental design.

1. The response variable is continuous.
2. The residuals follow the normal probability distribution with mean equal to zero and constant variance.
3. The subjects are independent.

Since in a within-subject design responses coming from the same subject are not independent, assumption 3 must be modified for responses within a subject. Independence between subjects is still assumed.

4. The within-subject covariance matrices are equal for all between-subject groups. In this type of experiment, the repeated measurements on a subject may be thought of as a multivariate response vector having a certain covariance structure. This assumption states that these covariance matrices are constant from group to group.
5. When using an *F* test, the within-subject covariance matrices are assumed to be *circular*. One way of defining circularity is that the variances of differences between any two measurements within a subject are constant for all measurements. Since responses that are close together in time often have a higher correlation than those that are far apart, it is common for this assumption to be violated. This assumption is not necessary for the validity of the three multivariate tests: Wilks' lambda, Pillai-Bartlett trace, or Hotelling-Lawley trace.

Advantages of Within-Subjects Designs

Because the response to stimuli usually varies less within an individual than between individuals, the within-subject variability is usually less than (or at most equal to) the between-subject variability. By reducing the underlying variability, the same power can be achieved with a smaller number of subjects.

Disadvantages of Within-Subjects Designs

1. *Practice effect*. In some experiments, subjects systematically improve as they practice the task being studies. In other cases, subjects may systematically get worse as the get fatigued or bored with the experimental task. Note that only the treatment administered first is immune to practice effects. Hence, experimenters should make an effort to balance the number of subjects receiving each treatment first.
2. *Carryover effect*. In many drug studies, it is important to wash out the influence of one drug completely before the next drug is administered. Otherwise, the influence of the first drug carries over into the response to the second drug.
3. *Statistical analysis.* The statistical model is more restrictive than in a regular factorial design since the individual responses must have certain mathematical properties.

Even in the face of all these disadvantages, repeated measures (within-subject) designs are popular in many areas of research. It is important that you recognize these problems going in so you can make sure that the design is appropriate, rather than learning of them later after the research has been conducted.

Technical Details

General Linear Multivariate Model

This section provides the technical details of the repeated measures designs that can be analyzed by ***PASS***. The approximate power calculations outlined in Muller, LaVange, Ramey, and Ramey (1992) are used. Using their notation, for N subjects, the usual general linear multivariate model is

$$\underset{(N\times p)}{Y} = \underset{(N\times q\times p)}{XM} + \underset{(N\times p)}{R}$$

where each row of the residual matrix R is distributed as a multivariate normal

$$row_k(R) \sim N_p(0,\Sigma)$$

Note that p is the product of the number of levels of each of the within-subject factors, q is the number of design variables, Y is the matrix of responses, X is the design matrix, M is the matrix of regression parameters (means), and R is the matrix of residuals.

Hypotheses about various sets of regression parameters are tested using

$$H_0\text{: } \underset{a\times b}{\Theta} = \Theta_0$$

$$\underset{a\times q\times p\times b}{CMD} = \Theta$$

where *C* and *D* are orthonormal contrast matrices and Θ_0 is a matrix of hypothesized values, usually zeros. Note that *C* defines contrasts among the between-subject factor levels and *D* defines contrast among the within-subject factor levels.

Tests of the various main effects and interactions may be constructed with suitable choices for *C* and *D*. These tests are based on

$$\hat{M} = \left(X'X\right)^{-} X'Y$$

$$\hat{\Theta} = C\hat{M}D$$

$$\underset{b\times b}{H} = \left(\hat{\Theta} - \Theta_0\right)' \left[C\left(X'X\right)^{-} C'\right]^{-1} \left(\hat{\Theta} - \Theta_0\right)$$

$$\underset{b\times b}{E} = D'\hat{\Sigma}D \cdot \left(N - r\right)$$

$$\underset{b\times b}{T} = H + E$$

where *r* is the rank of *X*.

Geisser-Greenhouse F Test

Upon the assumption that Σ has compound symmetry, a size α test of $H_0\!:\Theta = \Theta_0$ is given by the *F* ratio

$$F = \frac{\mathrm{tr}\left(H\right)/ab}{\mathrm{tr}\left(E\right)/\left[b\left(N-r\right)\right]}$$

with degrees of freedom given by

$$df1 = ab$$

$$df2 = b\left(N - r\right)$$

and noncentrality parameter

$$\lambda = df1\left(F\right)$$

The assumption that Σ has compound symmetry is usually not viable. Box (1954a,b) suggested that adjusting the degrees of freedom of the above *F*-ratio could compensate for the lack of compound symmetry in Σ. His adjustment has become known as the Geisser-Greenhouse adjustment. Under this adjustment, the modified degrees of freedom and noncentrality parameter are given by

$$df1 = ab\varepsilon$$

$$df2 = b\left(N - r\right)\varepsilon$$

$$\lambda = df1\left(F\right)\varepsilon$$

where

$$\varepsilon = \frac{\mathrm{tr}\left(D'\hat{\Sigma}D\right)^2}{b\,\mathrm{tr}\left(D'\hat{\Sigma}DD'\hat{\Sigma}D\right)}$$

The range of ε is $\frac{1}{b-1}$ to 1. When $\varepsilon = 1$, the matrix is *spherical*. When $\varepsilon = \frac{1}{b-1}$, the matrix differs maximally from sphericity.

Note that the Geisser-Greenhouse adjustment is only needed for testing main effects and interactions involving within-subject factors. Main effects and interactions that involve only between-subject factors need no such adjustment.

The critical value F_{Crit} is computed using the expected value of ε to adjust the degrees of freedom. That is, the degrees of freedom of F_{Crit} are given by

$$df1 = ab\mathrm{E}(\varepsilon)$$

$$df2 = b(N-r)\mathrm{E}(\varepsilon)$$

where

$$E(\hat{\varepsilon}) = \begin{cases} \varepsilon + \dfrac{g_1}{N-r} & \text{if } \varepsilon > \dfrac{g_1}{N-r} \\ \varepsilon / 2 & \text{otherwise} \end{cases}$$

$$g_1 = \sum_{i=1}^{T} f_{ii}\xi_i^2 + \sum\sum_{i \neq j} \frac{f_i \xi_i \xi_j}{(\xi_i - \xi_j)}$$

$$f_i = \frac{\partial \varepsilon}{\partial \xi_i}$$

$$= \frac{2\sum \xi_j}{df_1 \sum \xi_j^2} - \frac{2\lambda_i \left(\sum \xi_j\right)^2}{df_1 \left(\sum \xi_j^2\right)^2}$$

$$f_{ii} = \frac{\partial^{(2)} \varepsilon}{\partial \xi_i^{(2)}}$$

$$= 2h_1 - 8h_2 + 8h_3 - 2h_4$$

$$h_1 = \frac{2}{df_1 \sum \xi_j^2}$$

$$h_2 = \frac{\xi_i \left(\sum \xi_j\right)}{df_1 \left(\sum \xi_j^2\right)^2}$$

$$h_3 = \frac{\xi_i^2 \left(\sum \xi_j\right)^2}{df_1 \left(\sum \xi_j^2\right)^3}$$

$$h_4 = \frac{\left(\sum \xi_j\right)^2}{df_1 \left(\sum \xi_j^2\right)^2}$$

where the ξ_j's are the ordered eigenvalues of $D'\Sigma D$.

Wilks' Lambda Approximate F Test

The hypothesis $H_0:\Theta=\Theta_0$ may be tested using Wilks' likelihood ratio statistic *W*. This statistic is computed using

$$W=\left|ET^{-1}\right|$$

An *F* approximation to the distribution of *W* is given by

$$F_{df_1,df_2}=\frac{\eta/df_1}{(1-\eta)/df_2}$$

where

$$\lambda=df_1F_{df_1,df_2}$$

$$\eta=1-W^{1/g}$$

$$df1=ab$$

$$df2=g\left[(N-r)-(b-a+1)/2\right]-(ab-2)/2$$

$$g=\left(\frac{a^2b^2-4}{a^2+b^2-5}\right)^{\frac{1}{2}}$$

Pillai-Bartlett Trace Approximate F Test

The hypothesis $H_0:\Theta=\Theta_0$ may be tested using the Pillai-Bartlett Trace. This statistic is computed using

$$T_{PB}=tr\left(HT^{-1}\right)$$

A non-central *F* approximation to the distribution of T_{PB} is given by

$$F_{df_1,df_2}=\frac{\eta/df_1}{(1-\eta)/df_2}$$

where

$$\lambda=df_1F_{df_1,df_2}$$

$$\eta=\frac{T_{PB}}{s}$$

$$s=\min(a,b)$$

$$df1=ab$$

$$df2=s\left[(N-r)-b+s\right]$$

Hotelling-Lawley Trace Approximate F Test

The hypothesis $H_0 : \Theta = \Theta_0$ may be tested using the Hotelling-Lawley Trace. This statistic is computed using

$$T_{HL} = tr\left(HE^{-1}\right)$$

An *F* approximation to the distribution of T_{HL} is given by

$$F_{df_1,df_2} = \frac{\eta / df_1}{(1-\eta)/df_2}$$

where

$$\lambda = df_1 F_{df_1,df_2}$$

$$\eta = \frac{\frac{T_{HL}}{s}}{1+\frac{T_{HL}}{s}}$$

$$s = \min(a,b)$$

$$df1 = ab$$

$$df2 = s\left[(N-r)-b+s\right]$$

The M (Mean) Matrix

In the general linear multivariate model presented above, *M* represents a matrix of regression coefficients. Since you must provide the elements of *M*, we will discuss its meaning in more detail. Although other structures and interpretations of *M* are possible, in this module we assume that the elements of *M* are the cell means. The rows of *M* represent the between-subject categories and the columns of *M* represent the within-group categories.

The *q* rows of *M* represent the *q* groups into which the subjects can be classified. For example, if a design includes three between-subject factors with 2, 3, and 4 categories, the matrix *M* would have 2 *x* 3 *x* 4 = 24 rows. That is, *q* = 24. Similarly, if a design has three within-subject factors with 3, 3, and 3 categories, the matrix *M* would have 3 *x* 3 *x* 3 = 27 columns. That is, *p* = 27.

Consider now an example in which *q* = 3 and *p* = 4. That is, there are three groups into which subjects can be placed. Each subject is measured four times. The matrix *M* would appear as follows.

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{bmatrix}$$

For example, the element μ_{12} is the mean of the second measurement of subjects in the first group. To calculate the power of this design, you would need to specify appropriate values of all twelve means.

As a second example, consider a design with three between-subject factors and three within-subject factors, all of which have two categories. The *M* matrix for this design would be as follows.

$$M = \left[\begin{array}{ccc|cccccccc}
 & & W1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \\
 & & W2 & 1 & 1 & 2 & 2 & 1 & 1 & 2 & 2 \\
 & & W3 & 1 & 2 & 1 & 2 & 1 & 2 & 1 & 2 \\
B1 & B2 & B3 & & & & & & & & \\ \hline
1 & 1 & 1 & \mu_{111111} & \mu_{111112} & \mu_{111121} & \mu_{111122} & \mu_{111211} & \mu_{111212} & \mu_{111221} & \mu_{111222} \\
1 & 1 & 2 & \mu_{112111} & \mu_{112112} & \mu_{112121} & \mu_{112122} & \mu_{112211} & \mu_{112212} & \mu_{112221} & \mu_{112222} \\
1 & 2 & 1 & \mu_{121111} & \mu_{121112} & \mu_{121121} & \mu_{121122} & \mu_{121211} & \mu_{121212} & \mu_{121221} & \mu_{121222} \\
1 & 2 & 2 & \mu_{122111} & \mu_{122112} & \mu_{122121} & \mu_{122122} & \mu_{122211} & \mu_{122212} & \mu_{122221} & \mu_{122222} \\
2 & 1 & 1 & \mu_{211111} & \mu_{211112} & \mu_{211121} & \mu_{211122} & \mu_{211211} & \mu_{211212} & \mu_{211221} & \mu_{211222} \\
2 & 1 & 2 & \mu_{212111} & \mu_{212112} & \mu_{212121} & \mu_{212122} & \mu_{212211} & \mu_{212212} & \mu_{212221} & \mu_{212222} \\
2 & 2 & 1 & \mu_{221111} & \mu_{221112} & \mu_{221121} & \mu_{221122} & \mu_{221211} & \mu_{221212} & \mu_{221221} & \mu_{221222} \\
2 & 2 & 2 & \mu_{222111} & \mu_{222112} & \mu_{222121} & \mu_{222122} & \mu_{222211} & \mu_{222212} & \mu_{222221} & \mu_{222222}
\end{array} \right]$$

The subscripts for each mean follow the pattern $\mu_{B1\,B2\,B3\,W1\,W2\,W3}$. The first three subscripts indicate the between-subject categories and the second three subscripts indicate the within-subject categories. Notice that the first three subscripts are constant in each row and the second three subscripts are constant in each column.

Specifying the M Matrix

When computing the power in a repeated measures analysis of variance, the specification of the *M* matrix is one of your main tasks. The program cannot do this for you. The calculated power is directly related to your choice. So your choice for the elements of *M* must be selected carefully and thoughtfully. When authorization and approval from a government organization is sought, you should be prepared to defend your choice of *M*. In this section, we will explain how you can specify *M*.

Before we begin, it is important that you have in mind exactly what *M* is. *M* is a table of means that represent the size of the differences among the means that you want the study or experiment to detect. That is, *M* gives the means under the alternative hypothesis. Under the null hypothesis, these means are assumed to be equal. Because of the complexity of the repeated measures design, it is often difficult to choose reasonable values, so ***PASS*** will help you. But it is important to remember that you are responsible for these values and that the sample sizes calculated are based on them.

One way to specify the *M* matrix is to do so directly into the spreadsheet. You might do this if you are calculating the 'retrospective' power of a study that has already been completed, or if it is simply easier to write the matrix directly. Usually, however, you will specify the *M* matrix in portions.

We will begin our discussion of specifying the *M* matrix with an example. Consider a study of two groups of subjects. Each subject was tested, then a treatment was administered, then the subject was tested again at the ten minute mark, and then tested a third time after sixty minutes. The researchers wanted the sample size to be large enough to detect the following pattern in the means.

Table of Hypothesized Means				
	Time Period			
Group	T0	T10	T60	Average
A	100	130	100	110
B	120	180	120	140
Average	110	155	110	125

To understand how they derived this table, we will perform some basic arithmetic on it.

Step 1 – Remove the Overall Mean

Subtract 125, the overall mean, from each of the individual means.

Table of Hypothesized Means Adjusted for Overall Mean				
	Time Period			
Group	T0	T10	T60	Average
A	-25	5	-25	-15
B	-5	55	-5	15
Average	-15	30	-15	125

Step 2 – Remove the Group Effect

Subtract -15 from the first row and 15 from the second row.

Table of Hypothesized Means Adjusted for Group				
	Time Period			
Group	T0	T10	T60	Total
A	-10	20	-10	-15
B	-20	40	-20	15
Total	-15	30	-15	125

Step 3 – Remove the Time Effect

Subtract -15 from the first column, 30 from the second column, and -15 from the third column.

Table of Hypothesized Means

Adjusted for Group and Time

	Time Period				
Group	**T0**	**T10**	**T60**	**Effect**	**Effect + Overall**
A	5	-10	5	-15	110
B	-5	10	-5	15	140
Effect	-15	30	-15		
Effect + Overall	110	155	110		125

This table, called an effects table, lets us see the individual effect of each component of the model. For example, we can see that the hypothesized pattern across time is that T10 is 45 units higher than either endpoint. Similarly, we note that the hypothesized pattern for the two groups is that Group B is 30 units larger than Group A.

Understanding the interaction is more difficult. One interpretation focuses on T10. We note that in Group A the response for T10 is 10 less than expected while in Group B the response for T10 is 10 more than expected.

Entering This Information into *PASS*

Rather than enter the individual values into ***PASS***, you can enter the group, time, and interaction effects directly. For this example, you could enter '110 140' or '-15 15' for the hypothesized means of the between factor and '110 155 110' or '-15 30 -15' for the hypothesized means of the within factor. For the interaction, you would enter the six interaction values '5 -10 5 -5 10 -5'.

Another way to enter the interaction information would be to indicate that the size of the interaction to be detected is about half that of the group factor or about a third of the time factor. For a complete discussion of the interpretation of various interactions, we suggest that you look at Kirk (1982).

The C Matrix for Between-Subject Contrasts

The *C* matrix is comprised of contrasts that are applied to the rows of *M*. That is, these are between-group contrasts. You do not have to specify these contrasts. They are generated for you. You should understand that a different *C* matrix is generated for each between-subject term in the model. For example, in the six factor example above, the *C* matrix that will be generated for testing the between-subject factor B1 is

$$C_{B1} = \begin{bmatrix} \frac{-1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} \end{bmatrix}$$

Note that the divisor $\sqrt{8}$ is used so that the total of the squared elements is one. This is required so that the contrast matrix is *orthonormal*.

When creating a test for B1, the matrix D is created to average across all within-subject categories.

$$D_{B1} = \begin{bmatrix} \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} \end{bmatrix}$$

Generating the C Matrix when There are Multiple Between Factors

Generating the C matrix when there is more than one between factor is more difficult. We like the method of O'Brien and Kaiser (1985) which we briefly summarize here.

Step 1. Write a complete set of contrasts suitable for testing each factor separately. For example, if you have three factors with 2, 3, and 4 categories, you might use

$$\ddot{C}_{B1} = \begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},\ \ddot{C}_{B2} = \begin{bmatrix} \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},\text{ and } \ddot{C}_{B3} = \begin{bmatrix} \frac{-3}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} \\ 0 & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Step 2. Define appropriate J_k matrices corresponding to each factor. These matrices comprised of one row and k columns whose equal element is chosen so that the sum of its elements squared is one. In this example, we use

$$J_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},\ J_3 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix},\ J_4 = \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix}$$

Step 3. Create the appropriate contrast matrix using a direct (Kronecker) product of either the $\ddot{C}_{Bi}$ matrix if the factor is included in the term or the J_i matrix when the factor is not in the term. Remember that the direct product is formed by multiplying each element of the second matrix by all members of the first matrix. Here is an example

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 & 0 & 0 & -1 & -2 \\ 3 & 4 & 0 & 0 & -3 & -4 \\ 0 & 0 & 2 & 4 & 0 & 0 \\ 0 & 0 & 6 & 8 & 0 & 0 \\ -1 & -2 & 0 & 0 & 3 & 6 \\ -3 & -4 & 0 & 0 & 9 & 12 \end{bmatrix}$$

As an example, we will compute the *C* matrix suitable for testing factor *B*2

$$C_{B2} = J_2 \otimes \ddot{C}_{B2} \otimes J_4$$

Expanding the direct product results in

$$C_{B2} = J_2 \otimes \ddot{C}_{B2} \otimes J_4$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{-2}{\sqrt{12}} & \frac{-2}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} \\ 0 & 0 & \frac{-1}{\sqrt{4}} & \frac{-1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} \\ 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} \end{bmatrix}$$

Similarly, the *C* matrix suitable for testing interaction *B*2*B*3 is

$$C_{B2B3} = J_2 \otimes \ddot{C}_{B2} \otimes \ddot{C}_{B3}$$

We leave the expansion of this matrix *PASS*, but we think you have the idea.

The D Matrix for Within-Subject Contrasts

The *D* matrix is comprised of contrasts that are applied to the columns of *M*. That is, these are within-group contrasts. You do not have to specify these contrasts either. They will be generated for you. Specification of the *D* matrix is similar to the specification of the *C* matrix, except that now the matrices are all transposed.

Interactions of Between-Subject and Within-Subject Factors

Interactions that include both between-subject factors and within-subject factors require that between-subject portion be specified by the *C* matrix and the within-subject portion be specified with the *D* matrix.

Power Calculations

To calculate statistical power, we must determine distribution of the test statistic under the alternative hypothesis which specifies a different value for the regression parameter matrix *B*. The distribution theory in this case has not been worked out, so approximations must be used. We use the approximations given by Mueller and Barton (1989) and Muller, LaVange, Ramey, and Ramey (1992). These approximations state that under the alternative hypothesis, F_U is distributed as a noncentral *F* random variable with degrees of freedom and noncentrality shown above. The calculation of the power of a particular test may be summarized as follows

1. Specify values of X, M, Σ, C, D, and Θ_0.
2. Determine the critical value using $F_{crit} = FINV(1-\alpha, df1, df2)$, where *FINV*( ) is the inverse of the central *F* distribution and α is the significance level.
3. Compute the noncentrality parameter λ.
4. Compute the power as

$$Power = 1 - NCFPROB(F_{crit}, df1, df2, \lambda)$$

where *NCFPROB*( ) is the noncentral *F* distribution.

Covariance Matrix Assumptions

The following assumptions are made when using the *F* test. These assumptions are not needed when using one of the three multivariate tests.

In order to use the *F* ratio to test hypotheses, certain assumptions are made about the distribution of the residuals e_{ijk}. Specifically, it is assumed that the residuals for each subject, $e_{ij1}, e_{ij2}, \cdots, e_{ijT}$, are distributed as a multivariate normal with means equal to zero and covariance matrix Σ_{ij}. Two additional assumptions are made about these covariance matrices. First, they are assumed to be equal for all subjects. That is, it is assumed that $\Sigma_{11} = \Sigma_{12} = \cdots = \Sigma_{Gn} = \Sigma$. Second, the covariance matrix is assumed to have a particular form called *circularity*. A covariance matrix is *circular* if there exists a matrix *A* such that

$$\Sigma = A + A' + \lambda I_T$$

where I_T is the identity matrix of order *T* and λ is a constant.

This property may also be defined as

$$\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij} = 2\lambda$$

One type of matrix that is circular is one that has *compound symmetry*. A matrix with this property has all elements on the main diagonal equal and all elements off the main diagonal equal. An example of a covariance matrix with compound symmetry is

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix}$$

or, with actual numbers,

$$\begin{bmatrix} 9 & 2 & 2 & 2 \\ 2 & 9 & 2 & 2 \\ 2 & 2 & 9 & 2 \\ 2 & 2 & 2 & 9 \end{bmatrix}$$

An example of a matrix which does not have compound symmetry but is still circular is

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 3 & 4 & 5 \\ 3 & 6 & 5 & 6 \\ 4 & 5 & 8 & 7 \\ 5 & 6 & 7 & 10 \end{bmatrix}$$

Needless to say, the need to have the covariance matrix circular is a very restrictive assumption.

Between-Subject Standard Deviation

The subject-to-subject variability is represented by $\sigma^2_{Between}$. In the repeated measures AOV table, this quantity is estimated by the between subjects mean square (*MSB*). This quantity is calculated from Σ using the formula

$$\sigma^2_{Between} = \frac{\sum_{i=1}^{T}\sum_{j=1}^{T}\sigma_{ij}}{T}$$
$$= \frac{\sum_{i=1}^{T}\sum_{j=1}^{T}\sigma_{ii}\sigma_{jj}\rho_{ij}}{T}$$

When Σ has compound symmetry, which requires all $\sigma_{ii} = \sigma$ and all $\rho_{ij} = \rho$, the above formula reduces to

$$\sigma^2_{Between} = \sigma^2\left(1 + (T-1)\rho\right)$$

Note that *F* tests of between factors and their interactions do not require the circularity assumption so the Geisser-Greenhouse correction is not applied to these tests.

Within-Subject Standard Deviation

The within-subject variability is represented by σ^2_{Within}. In the repeated measures AOV table, this quantity is estimated by the within-subjects mean square (*MSW*). This quantity is calculated from Σ using the formula

$$\sigma^2_{Within} = \frac{\sum_{i=1}^{T}\sigma_{ii}}{T} - \frac{2\sum_{i=1}^{T}\sum_{j=i+1}^{T}\sigma_{ij}}{T(T-1)}$$

$$= \frac{\sum_{i=1}^{T}\sigma_{ii}}{T} - \frac{2\sum_{i=1}^{T}\sum_{j=i+1}^{T}\rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}}{T(T-1)}$$

When Σ has compound symmetry, which requires all $\sigma_{ii} = \sigma$ and all $\rho_{ij} = \rho$, the above formula reduces to

$$\sigma^2_{Within} = \sigma^2(1-\rho)$$

Estimating Sigma and Rho from Existing Data

Using the above results for existing data, approximate values for σ and ρ may be estimated from a previous analysis of variance table that provides estimates of MSB and MSW. Solving the above equations for σ and ρ yields

$$\rho = \frac{\sigma^2_{Between} - \sigma^2_{Within}}{\sigma^2_{Between} + (T-1)\sigma^2_{Within}}$$

$$\sigma^2 = \frac{\sigma^2_{Within}}{1-\rho}$$

Substituting MSB for $\sigma^2_{Between}$ and MSW for σ^2_{Within} yields the estimates

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (T-1)MSW}$$

$$\hat{\sigma}^2 = \frac{MSW}{1-\hat{\rho}}$$

Note that these estimators assume that the design meets the circularity assumption, which is usually not the case. However, they provide crude estimates that can be used in planning.

Procedure Options

This section describes the options that are unique to this procedure. To find out more about using the other tabs, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for. If you choose to solve for *n* (sample size), you must also specify which test statistic you want to use.

When you choose to solve for *n*, the program searches for the lowest sample size that meets the alpha and power criterion you have specified for each of the terms. If you do not want a term to be used in the search, set its levels to empty or 0.

Also, when the '= n's' box is not checked, the search is made using unequal group sample sizes. The relative proportion of the sample in each group is set by the values of *n* given in the Subjects Per Group box. For example, if your design has three groups and you entered '1 1 2' in the Subjects Per Group box, the search will only consider designs in which the size of the last group is twice the rest. That is, it will consider '2 2 4', '3 3 6', '4 4 8', etc.

Note: no plots are generated when you solve for *n*.

Sample Size

n (Subjects Per Group)

Specify one or more values for the number of subjects per group. The total sample size is the sum of the individual group sizes across all groups.

You can specify a list like '2 4 6'. The items in the list may be separated with commas or blanks. The interpretation of the list depends on the =n's check box. When the =n's box is checked, a separate analysis is calculated for each value of *n*. When the =n's box is not checked, ***PASS*** uses the *n's* as the actual group sizes. In this case, the number of items entered must match the number of groups in the design.

When you choose to solve for *n* and the '= n's' box is not checked, the search is made using unequal group sample sizes. The relative proportion of the sample in each group is set by the values of *n* given in this box. For example, if your design has three groups and you enter '1 1 2' here, the search will only consider designs in which the size of the last group is twice the rest. That is, it will consider '2 2 4', '3 3 6', '4 4 8', etc.

= n's

This option controls whether the number of subjects per group is to be equal for all groups or not. When checked, the number of subjects per group is equal for all groups. A list of values such as '5 10 15' represents three designs: one with five per group, one with ten per group, and one with fifteen per group.

When this option is not checked, the n's are assumed to be unequal. A list of values represents the size of the individual groups. For example, '5 10 15' represents a single, three-group design with five in the first group, ten in the second group, and fifteen in the third group.

Effect Size – Means

Means Matrix

Use this option to the specify spreadsheet columns containing a hypothesized means matrix that will be used to compute the Sm values. All individual Sm values are ignored. You can obtain the spreadsheet by selecting 'Window', then 'Data', from the menus.

The between factors are represented across the columns of the spreadsheet and the within factors are represented down the rows. The number of columns specified must equal the number of groups. The number of rows with data in these columns must equal the number of times. For example, suppose you are designing an experiment that is to have two between factors (A & B) and two within factors (D & E). Suppose each of the four factors has two levels. The columns of the spreadsheet would be

A1B1 A1B2 A2B1 A2B2.

The rows of the spreadsheet would represent

D1E1

D1E2

D2E1

D2E2

Example

To see how this option works, consider the following table of hypothesized means for an experiment with one between factor (A) having two groups and one within factor (B) having three time periods. The values in columns C1 and C2 of the spreadsheet are

C1	C2
2.0	4.0
4.0	6.0
6.0	11.0

By subtracting the appropriate means, the following table of effects results

```
             C1       C2       | Means| Effects
Row1          0.5     -0.5     | 3.0   | -2.5
Row2          0.5     -0.5     | 5.0   | -0.5
Row3         -1.0      1.0     | 8.5   |  3.0
             ---      ---      | ---
Means         4.0      7.0     | 5.5
Effects      -1.5      1.5
```

The standard deviation of the A effects is calculated as

$$\begin{aligned}\sigma_A &= \sqrt{\frac{(-1.5)^2 + (1.5)^2}{2}} \\ &= \sqrt{2.25} \\ &= 1.5\end{aligned}$$

The standard deviation of the B effects is calculated as

$$\begin{aligned}\sigma_B &= \sqrt{\frac{(-2.5)^2 + (-0.5)^2 + (3.0)^2}{3}} \\ &= \sqrt{\frac{15.5}{3}} \\ &= 2.27\end{aligned}$$

The standard deviation of the interaction effects is found to be

$$\begin{aligned}\sigma_{AB} &= \sqrt{\frac{(0.5)^2 + (0.5)^2 + (-1.0)^2 + (-0.5)^2 + (-0.5)^2 + (1.0)^2}{6}} \\ &= \sqrt{\frac{3.0}{6}} \\ &= 0.71\end{aligned}$$

These three standard deviations are used to represent the effect sizes of the corresponding terms.

Discussion

When using this option, it is less confusing to concentrate on a single term at a time. For example, consider a 2-by-4 design in which your primary interest is in testing the AB interaction. Instead of trying to determine a means matrix the will represent factor A, factor B, and the AB interaction, ignore factor A and factor B and just consider the interaction. You might want to consider the following pattern

C1	C2
0.0	0.0
0.0	1.0
0.0	2.0
0.0	3.0

That is, the first group remains constant while the second group increases for 0.0 to 3.0.

By specifying various values for K (the means multiplier), you can study to impact of increasing the values. For example, when K is set to 2, the above means matrix becomes

C1	C2
0.0	0.0
0.0	2.0
0.0	4.0
0.0	6.0

Thus, by simply changing K, several scenarios may be studied. (We wish to thank Keith Muller for suggesting this method of specifying the Sm values.)

K (Means Multipliers)

These values are multiplied times the means matrix to give you various effect sizes. A separate power calculation is generated for each value of K. These values become the horizontal axis in the second power chart. For example, if an Sm value is 80, setting this option to ‘50 100 150’ would result in three Sm values: 40, 80, and 120. If you want to ignore this setting, enter ‘1’.

Note that when you enter Sm values directly, ***PASS*** generates an appropriate means matrix and then multiplies this matrix by each of these K values.

Effect Size – Between- and Within-Subject Factors

Label

Specify a label for this factor. Although we suggest that only a single letter be used, the label can consist of several letters. When several letters are used, the labels for the interactions may be extra long and confusing. Of course, you must be careful not to use the same label for two factors.

One of the easiest sets of labels is to use A, B, and C for the between factors and D, E, and F for the within factors. A useful alternative is to use B1, B2, and B3 for the between factors and W1, W2, and W3 for the within factors.

Levels

Specify the number of levels (categories) in this factor. Typical values are from 2 to 8. Set this to a blank (or 0) to ignore the factor in the design.

Alpha

These options specify the probability of a type-I error (alpha) for each factor and interaction. A type-I error occurs when you reject the null hypothesis of zero effects when in fact they are zero. Since they are probabilities, alpha values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This value may be interpreted as meaning that about one F-test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You can specify different alpha values for different terms. For example, although you have three terms in an experiment, you might be mainly interested in only one of them. Hence, you could increase the alpha level of the tests of the other terms and thereby increase their power. Also, you may want to increase the alpha level of the interaction terms, as these will often have poor power otherwise.

Power or Beta

These options specify the power or beta (depending on the chosen setting) for each factor and interaction. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when you fail to reject the null hypothesis of equal effects when in fact they are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

Sm (Standard Deviation of Effects)

Enter the standard deviation of the effects ($\sigma_{effects}$) for this factor or interaction. This value represents the magnitude of the differences among the means (effects) that is to be detected.

The value of Sm may be entered in several ways: directly, as a list of numbers, or as a percentage of another term.

- **Directly**

  You can enter the value of Sm directly by specifying a single number. If only a single number is entered, it becomes the value of Sm.

  You can use the Standard Deviation Estimator window to calculate the value of Sm for various sets of means. This window is obtained by selecting PASS, then Other, and then Standard Deviation Estimator from the menus.

- **List of Numbers**

  When a list of numbers is entered, the standard deviation of those numbers is computed and used as the value of Sm. The numbers in the list may represent means or effects. The list may be a simple list, the STEP command, or the RANGE command.

  Simple List

  A Simple List is a set of numbers separated by blanks or commas. Examples a simple list are

  5 20 60

  2,5,7

  -4,0,6,9

  STEP Command

  The syntax of the STEP command is *STEP Start Inc*. The list begins at *Start* and increases by *Inc*. The number of values generated is determined by the number of levels in the term. Examples of the STEP command for a term with four levels are

  ‘STEP 0 2’ results in ‘0 2 4 6’.

  ‘STEP 1 -1’ results in ‘1 0 -1 -2’.

  ‘STEP 1 0.5’ results in ‘1 1.5 2 2.5’.

  RANGE Command

  The syntax of the RANGE command is *RANGE Minimum Maximum*. The list of numbers generated increase steadily from *Minimum* to *Maximum*. The RANGE command is handy when you want to vary the number of levels while keeping the values in a known range. Examples of the RANGE command for a term with four levels are

  ‘RANGE 10 70’ results in ‘10 30 50 70’.

  ‘RANGE 0 1’ results in ‘0 0.33 0.67 1’.

  ‘RANGE 1 4’ results in ‘1 2 3 4’.

Percentage of another Term

You can specify the value as a percentage of another term. The syntax of this command is *TERM PCT* where *TERM* is any other main effect or interaction in the model and *PCT* is a percentage. This method is often used to specify interaction Sm's. Examples of the PERCENTAGE command are

'A 100' results in an Sm equal to the Sm of factor A.

'B 50' results in an Sm equal to one-half of the Sm of factor B.

Interactions Tab

The values of Alpha, Power, and Sm are entered for various groups of 2-way and higher-order interactions.

Covariance Tab

This tab specifies the covariance matrix.

Covariance Matrix Specification

Specify Which Covariance Matrix Input Method to Use

This option specifies which method will be used to define the covariance matrix.

- **Standard Deviations and Autocorrelations**

  This option generates a covariance matrix based on the settings for the standard deviations (SD's) and the pattern of autocorrelations as specified in the options on this screen down to and including 'R2'. More about this option is given below.

- **Covariance Matrix Variables**

  When this option is selected, the covariance matrix is read in from the columns of the spreadsheet. This is the most flexible method, but specifying a covariance matrix is tedious. You will usually only use this method when a specific covariance is given to you. More about this option is given below.

  Note that the spreadsheet is shown by selecting the menus: 'Window' and then 'Data'.

Time Metric

This option is used when the 'Specify How the Standard Deviations Change Across Time' option is set to 'Range from SD1 to SD2 using the Time Metric' to help define the covariance matrix. It specifies a sequential list of time points at which measurements of the subjects are made. Often, measurements are made at equally-spaced points through time. This is not always the case. It is important to define a time metric that corresponds to the study. For example, measurements might be planned at the beginning, after one day, after one week, and after one month.

The number of time points is the product of the number of levels of all within factors.

The time metric influences the values of the SD's as well as the correlations between two measurements on the same individual.

Entering a List of Times

A list of times can be entered in which the time values are separated by blanks or commas. The time metric can be defined in any time scale desired. For example, you could enter 0, 0.143, 1, 2, 3 if times were 0 weeks, 1/7 week (day 1), 1 week, 2 weeks, 3 weeks. The same values in days would be 0, 1, 7, 14, 21.

Using the RANGE Command

The RANGE command can be used to specify a list of times. The syntax of the RANGE command is

```
RANGE Minimum Maximum
```

A set of equal-spaced time points is generated between Minimum and Maximum. The number of time points depends on the number of within-factor levels.

This setting is very useful when you want to study the impact of increasing/decreasing the number of measurements per subject during the same period of time. That is, if the study will last five weeks, will the power of the statistical tests increase if you take ten measurements rather than five?

For example, suppose there are six times. Entering

```
RANGE 0 10
```

will generate the time metric: 0, 2, 4, 6, 8, 10. If the number of times is changed to eleven, the time metric will become: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Using the STEP Command

The STEP command can be used to specify a list of times. The syntax of the STEP command is

```
STEP Start, Inc
```

This command generates time points beginning at *Start* and incrementing by *Inc*. For example,

```
STEP 0 2
```

would generate the values 0, 2, 4, 6, 8, …

Covariance Matrix Specification – Within-Subject Standard Deviation Pattern

The parameters in this section provide a flexible way to specify Σ , the covariance matrix. Because the covariance matrix is symmetric, it can be represented as

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1T} & \sigma_{2T} & \cdots & \sigma_{TT} \end{bmatrix}$$

$$
= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_T\rho_{1T} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \cdots & \sigma_2\sigma_T\rho_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_T\rho_{1T} & \sigma_2\sigma_T\rho_{2T} & \cdots & \sigma_T^2 \end{bmatrix}
$$

$$
= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1T} \\ \rho_{12} & 1 & \cdots & \rho_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1T} & \rho_{2T} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T \end{bmatrix}
$$

where T is the product of the number of levels of all of the within factors.

Thus, the covariance matrix can be represented with complete generality by specifying the standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_T$ and the correlation matrix

$$
R = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1T} \\ \rho_{12} & 1 & \cdots & \rho_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1T} & \rho_{2T} & \cdots & 1 \end{bmatrix}.
$$

Specify How the Standard Deviations Change Across Time

This option specifies the method used to specify the standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_T$. Based on the method selected, the actual values are specified using SD1 and, in some cases, SD2.

Each σ is an estimate of the standard deviation that occurs when the same individual is measured at the same point in time under identical treatment conditions. It is a measure of the within-subject variability.

The available options are

- **Constant (Use SD1. Ignore SD2)**

  When this option is selected, the standard deviations are assumed to be equal. That is, it is assumed that $\sigma_1 = \cdots = \sigma_T$. The value of σ_i is specified in the SD1 field. The value in the SD2 field is ignored.

- **List of Standard Deviations (Use list in SD1. Ignore SD2.)**

  When this option is selected, a list of standard deviations can be entered in SD1. The items in the list can be separated by commas or blanks. The first value in the list becomes σ_1, the second value becomes σ_2, and so on. If the number of values in the list is less than the number of standard deviations required, the last value in the list is repeated. Note that all standard deviations in the list must be positive numbers.

- **Range from SD1 to SD2 using the Time Metric**

  When this option is selected, the standard deviations are spread between σ_1 and σ_T according to the spread in the Time Metric. The value in SD1 becomes σ_1 and the value in SD2 becomes σ_T .

  For example, suppose SD1 = 100, SD2 = 200, and the Time Metric values are 0, 2, 4, 10. The standard deviations would be

  $$\sigma_1 = 100.0 \qquad \sigma_2 = 120.0 \qquad \sigma_3 = 140.0 \qquad \sigma_4 = 200.0$$

SD1 (Standard Deviation 1)

This option is used to generate the covariance matrix. Its interpretation depends on the 'How the SD's Change Across Time' option's setting. Fundamentally, this is the standard deviation that occurs when the same individual is measured at the same point in time under identical treatment conditions. It is a measure of the within-subject variability.

You may want to use the special window that has been prepared to estimate SD1 from the mean square between (MSB) and the mean square within (MSW) of an existing table. To display this special window, from the menus select 'PASS', then 'Other', and then 'Standard Deviation Estimator'. Click on the 'Covariance Matrix' tab. Enter the values from the ANOVA table. The resulting value of 'Sigma' should be placed here.

Specify How the Standard Deviations Change Across Time = Constant

The value entered here is used as the standard deviation for all time points.

Specify How the Standard Deviations Change Across Time = List of Standard Deviations

The values in the list entered here become the values of the standard deviations. If the number in the list is less than the number required, the last value in the list is repeated.

Specify How the Standard Deviations Change Across Time = Range from SD1 to SD2 using the Time Metric

The value entered here is used as a beginning standard deviation, the value in SD2 is used as an ending standard deviation, and the intermediate standard deviations are spaced between SD1 and SD2 proportional to the values of the Time Metric.

For example, suppose SD1 = 100, SD2 = 200, and the Time Metric values are 0,2,4,10. The standard deviation values would be:

S(1)=100

S(2)=120

S(3)=140

S(4)=200

SD2 (Standard Deviation 2)

This parameter is used when 'How the SD's Change Across Time' option is set to 'Range…'. In that case, this option specifies the value of σ_T .

Covariance Matrix Specification – Autocorrelation Pattern

Specify How the Autocorrelations Change Across Time

This option specifies the pattern of the autocorrelations in the variance-covariance matrix. Three options are possible:

- **Constant**

  The value of R1 is used as the constant autocorrelation until the maximum time difference is reached, then the value of R2 is used. For example, if the maximum time difference is 3, R1 = 0.6, R2 = 0.1, and *T* = 6, the correlation matrix would appear as

$$R=\begin{bmatrix} 1 & 0.600 & 0.600 & 0.600 & 0.100 & 0.100 \\ 0.600 & 1 & 0.600 & 0.600 & 0.600 & 0.100 \\ 0.600 & 0.600 & 1 & 0.600 & 0.600 & 0.600 \\ 0.600 & 0.600 & 0.600 & 1 & 0.600 & 0.600 \\ 0.100 & 0.600 & 0.600 & 0.600 & 1 & 0.600 \\ 0.100 & 0.100 & 0.600 & 0.600 & 0.600 & 1 \end{bmatrix}$$

  Note that when all correlations are equal, this is the correlation pattern that is assumed by the repeated measure ANOVA F-test. It may be a good first approximation, but many researchers believe the next option (first-order autocorrelation) is more realistic.

- **1st Order Autocorrelation**

  The value of R1 is used as the base autocorrelation in a first-order, serial correlation pattern. For example, if the maximum time difference is 3, R1 = 0.6, R2 = 0.1, and *T* = 6, the correlation matrix would appear as

$$R=\begin{bmatrix} 1 & 0.600 & 0.360 & 0.216 & 0.100 & 0.100 \\ 0.600 & 1 & 0.600 & 0.360 & 0.216 & 0.100 \\ 0.360 & 0.600 & 1 & 0.600 & 0.360 & 0.216 \\ 0.216 & 0.360 & 0.600 & 1 & 0.600 & 0.360 \\ 0.100 & 0.216 & 0.360 & 0.600 & 1 & 0.600 \\ 0.100 & 0.100 & 0.216 & 0.360 & 0.600 & 1 \end{bmatrix}$$

  This pattern is often chosen as the most realistic when little is known about the correlation pattern.

- **Custom**

  The values of R1, R2, A, V, and Max Time Diff are used to generate a custom autocorrelation pattern. This relationship is modeled using the equation

$$\begin{aligned} Corr\left(Y_{si},Y_{sj}\right) &= \rho_{ij} \\ &= d\left(R1^{1-A+A\left|t_i-t_j\right|^V}\right)+\left(1-d\right)R2 \end{aligned}$$

where R1 is the base correlation, t_i and t_j are two time points, and A and V are specified constants. The variable *d* is one if $|t_i - t_j|$is less than Max Time Diff and zero otherwise.

Machin, Campbell, Fayers, and Pinol (1997) state that values of R1 between 0.60 and 0.75 are common.

We will present some examples to show you how this formula may be interpreted. For the moment, assume that the time metric is four, equally space time points of 1, 2, 3, and 4. Also, assume that Max Time Diff is set to 20.

Example 1

Let *A* = 0, *V* = 1, and $\rho_1 = \rho$. The correlation matrix becomes

$$R = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

Example 2

Let *A* = 1, *V* = 1, and $\rho_1 = \rho$. The correlation matrix becomes

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

which is the first-order autoregression model, a popular model in time series analysis.

Example 3

Let *A* = 1, *V* = 2, and $\rho_1 = \rho$. The correlation matrix becomes

$$R = \begin{bmatrix} 1 & \rho^2 & \rho^4 & \rho^6 \\ \rho^2 & 1 & \rho^2 & \rho^2 \\ \rho^4 & \rho^2 & 1 & \rho^2 \\ \rho^6 & \rho^4 & \rho^2 & 1 \end{bmatrix}$$

which is similar to Example 2 except that the correlations die out much more quickly.

Example 4

Let A = 0.5, V = 1, and $\rho_1 = \rho$. The correlation matrix becomes

$$R = \begin{bmatrix} 1 & \rho^{0.5} & \rho^{1.0} & \rho^{1.5} \\ \rho^{0.5} & 1 & \rho^{0.5} & \rho^{1.0} \\ \rho^{1.0} & \rho^{0.5} & 1 & \rho^{0.5} \\ \rho^{1.5} & \rho^{1.0} & \rho^{0.5} & 1 \end{bmatrix}$$

which is similar to Example 2 except that the correlations die out much more slowly.

Example 5

Let A = 1 and V = 1. For this example, set the Max Time Diff option to 2. The correlation matrix becomes

$$R = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_2 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Notice that this scenario lets you create a banded correlation matrix with two unique correlations.

Example 6

This example shows how this formula works when the Max Time Diff is set to 7 and the time metric is 1, 2, 7, 15. Let A = 1 and V = 1. The correlation matrix becomes

$$R = \begin{bmatrix} 1 & \rho_1 & \rho_1^6 & \rho_2 \\ \rho_1 & 1 & \rho_1^5 & \rho_2 \\ \rho_1^6 & \rho_1^5 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 \end{bmatrix}$$

R1 (Autocorrelation)

This is the autocorrelation, r1, between two measurements made on a subject at two time points that differ by one time unit. This value is combined with the other parameters in this section to form the covariance matrix.

Since this is a type of correlation, possible values range from -1 to 1. However, in this situation, a positive value is usually assumed, so the range is 0 to 1. A value near 0 indicates low autocorrelation. A value near 1 indicates high autocorrelation.

The value of this parameter depends on the Time Metric that is defined. Normally, you would expect a larger autocorrelation if the time metric units were in hours rather than days. In their book on sample size, Machin and Campbell comment the values between 0.60 and 0.75 are typical.

It is reasonable to assume that there is a correlation between two measurements made on the same subject at two points in time. It is often reasonable to assume that the size of this correlation diminishes as the two time points are further and further apart. That is, you would expect a much larger autocorrelation between two measurements taken one minute apart than between two measurements taken one week apart.

You may want to use the special window that has been prepared to estimate R1 from the mean square between (MSB) and the mean square within (MSW) of an existing table. To display this special window, from the menus select ‘PASS’, then ‘Other’, and then ‘Standard Deviation Estimator’. Click on the ‘Covariance Matrix’ tab. Enter the values from the ANOVA table. The resulting value of ‘Rho should be placed here.

R2 (Second AC)

This is the value of the secondary autocorrelation, R2. This value is used when the difference between two time points (see Time Metric) is greater than the value of Max Time Diff. Hence, if you set Max Time Diff to zero, this value will be used to calculate all correlations in the covariance matrix. When used, think of R2 as the correlation between measurements made on the same subject, regardless Of how far apart in time they are. Since we are assuming a positive autocorrelation, this value ranges between 0 and 1.

Max Time Diff

This is the maximum time difference (MTD) between two measurement points before the autocorrelation is set to the constant correlation value, R2.

In the autocorrelation formula:

$$\rho_{i,j} = dR1^{|t_i - t_j|^V} + (1-d)R2$$

The parameter *d* is equal to 1 when $|t_i - t_j| <$ MTD and 0 otherwise. Hence, this value controls when the autocorrelation is set to R2.

If you think of R2 as the correlation between measurements made on the same subject, regardless of how far apart in time they are, then this value should be set to that time difference at which the measurements times are no longer a factor.

For example, you might postulate that the autocorrelations are 0.9, 0.5, 0.2, 0.2, 0.2, and so on. That is, the autocorrelation is 0.9 for measurements taken one day apart, 0.5 for measurements taken two days apart, and 0.2 for all others. In this case, you would set Max Time Diff to 3.

A, V

These parameters are used when a 'Custom' autocorrelation pattern across time is specified. The formula used to calculate the autocorrelation is:

$$\rho_{i,j} = dR1^{1-A+A|t_i - t_j|^V} + (1-d)R2$$

where *d* is 1 if the time difference is <= Max Time Diff and 0 otherwise.

A = 0, V = 1 gives constant autocorrelation.

A = 1, V = 1 gives first-order autocorrelation.

Usually, V is set to 1. V should only be set to 1 or 2. The value of 1 is recommended. Set V at 2 when you want the autocorrelations to taper off rapidly as time occurs between measurements.

Covariance Matrix Specification – Covariance Matrix Variables

This option instructs the program to read the covariance matrix from the spreadsheet.

Spreadsheet Columns Containing the Covariance Matrix

This option designates the columns on the current spreadsheet holding the covariance matrix. It is used when the 'Specify Which Covariance Matrix Input Method to Use' option is set to *Covariance Matrix Variables*.

The number of columns and number of rows must match the number of time periods at which the subjects are measured.

Reports Tab

This tab specifies which reports and graphs are displayed as well as their format.

Select Output – Numeric Reports

Test in Summary Statement(s)

Indicate the test that is to be reported on in the Summary Statements.

Select Output – Report Options

Maximum Term-Order Reported

Indicate the maximum order of terms to be reported on. Occasionally, higher-order interactions are of little interest and so they may be omitted. For example, enter a '2' to limit output to individual factors and two-way interactions.

Skip Line After

The names of the terms can be too long to fit in the space provided. If the name contains more characters than this, the rest of the output is placed on a separate line. Enter '1' when you want every term's results printed on two lines. Enter '100' when you want every variable's results printed on one line.

Example 1 – Determining Sample Size

Researchers are planning a study of the impact of a drug on heart rate. They want to evaluate the differences in heart rate among three age groups: 20-40, 41-60, and over 60. Their experimental protocol calls for a baseline heart rate measurement, followed by administration of a certain level of the drug, followed by three additional measurements 30 minutes apart. They want to be able to detect a 10% difference in heart rate among the age groups. They want to detect 5% difference in heart rate within an individual across time. They decide the experiment should detect interaction effects of the same magnitude as the time factor. They plan to analyze the data using a Geisser-Greenhouse corrected F test.

Similar studies have found an average heart rate of 93, a standard deviation of 4, and an autocorrelation between adjacent measurements on the same individual of 0.7. The researchers assume that first-order autocorrelation adequately represents the autocorrelation pattern.

From a heart rate of 93, a 10% reduction gives 84. They decide on the age-group means of 93, 87, and 84. Similarly, a 5% reduction within a subject would result in a heart rate of 88. They decide on time means of 93, 89, 88, and 91.

How many subjects per age group are needed to achieve 95% power and a 0.05 significance level?

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**2 to 8 by 1**
=n's	**checked**
Means Matrix	blank
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**B**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**93 87 84**
For First Within-Subject Factor	
Label	**W**
Levels	**4**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**93 89 88 91**
Interactions Tab	
2-Way(Mixed) Interaction	
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**W 100**
Covariance Tab	
Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**4**
Time Metric	*Ignored*
Specify How Autocorr's Change	**1st Order Autocorr**
R1 (Autocorrelation)	**0.7**
Max Time Diff.	**100** (This large value will cause R2 to be ignored.)

Reports Tab

Numeric Results by Term	**Checked**
Numeric Results by Design	**Checked**
Regular F Test	**Not checked**
GG F Test	**Checked**
Wilks' Lambda	**Not checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
GG Detail Report	**Checked**
Means Matrix	**Checked**
Covariance Matrix	**Checked**
Test in Summary Statement	**GG F Test**
Show Plot 1	**Checked**
Show Plot 2	**Not checked**
Test That is Plotted	**GG F Test**
Max Term-Order Plotted	**2**
Max Term-Order Reported	**2**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Design Report

```
                                            Multiply   SD of  Standard
                                               Means Effects Deviation  Effect
Term    Test      Power      n       N          By    (Sm)   (Sigma)    Size     Alpha       Beta
B(3)    GG F      0.3096     2       6        1.00    3.74      3.29    1.14    0.0500     0.6904
W(4)    GG F      0.3266     2       6        1.00    1.92      1.31    1.47    0.0500     0.6734
BW      GG F      0.1588     2       6        1.00    1.92      1.31    1.47    0.0500     0.8412

B(3)    GG F      0.6459     3       9        1.00    3.74      3.29    1.14    0.0500     0.3541
W(4)    GG F      0.8099     3       9        1.00    1.92      1.31    1.47    0.0500     0.1901
BW      GG F      0.6267     3       9        1.00    1.92      1.31    1.47    0.0500     0.3733

B(3)    GG F      0.8478     4      12        1.00    3.74      3.29    1.14    0.0500     0.1522
W(4)    GG F      0.9486     4      12        1.00    1.92      1.31    1.47    0.0500     0.0514
BW      GG F      0.8598     4      12        1.00    1.92      1.31    1.47    0.0500     0.1402

B(3)    GG F      0.9415     5      15        1.00    3.74      3.29    1.14    0.0500     0.0585
W(4)    GG F      0.9871     5      15        1.00    1.92      1.31    1.47    0.0500     0.0129
BW      GG F      0.9535     5      15        1.00    1.92      1.31    1.47    0.0500     0.0465

B(3)    GG F      0.9793     6      18        1.00    3.74      3.29    1.14    0.0500     0.0207
W(4)    GG F      0.9970     6      18        1.00    1.92      1.31    1.47    0.0500     0.0030
BW      GG F      0.9860     6      18        1.00    1.92      1.31    1.47    0.0500     0.0140

B(3)    GG F      0.9931     7      21        1.00    3.74      3.29    1.14    0.0500     0.0069
W(4)    GG F      0.9993     7      21        1.00    1.92      1.31    1.47    0.0500     0.0007
BW      GG F      0.9961     7      21        1.00    1.92      1.31    1.47    0.0500     0.0039

B(3)    GG F      0.9978     8      24        1.00    3.74      3.29    1.14    0.0500     0.0022
W(4)    GG F      0.9999     8      24        1.00    1.92      1.31    1.47    0.0500     0.0001
BW      GG F      0.9990     8      24        1.00    1.92      1.31    1.47    0.0500     0.0010
```

The *Design Report* gives the power for each term in the design for each value of *n*. It is useful when you want to compare the powers of the terms in the design at a specific sample size.

In this example, the design goals of 0.95 power on all terms are achieved for *n* = 6.

The definitions of each of the columns of the report are as follows.

Term

This column contains the identifying label of the term. The number of levels for a factor is given in parentheses.

Test

This column identifies the test statistic. Since the power depends on the test statistic, you should make sure that this is the test statistic that you will use.

Power

This is the computed power for the term.

n

The value of *n* is the number of subjects per group.

N

The value of *N* is the total number of subjects in the study.

Multiply Means By

This is the value of the means multiplier, *K*.

SD of Effects (Sm)

This is the standard deviation of the effects σ_m for this term.

Standard Deviation

This is the value of σ, the random variation that σ_m is compared against by the *F* test. See the Technical Details for details on how these values are calculated.

Effect Size

The Effect Size is calculated by the expression σ_m / σ. It is an index of the size of the effect values relative to the standard deviation. Its value may be compared from experiment to experiment, regardless of the scale of the response variable.

Alpha

Alpha is the significance level of the test

Beta

Beta is the probability of failing to reject the null hypothesis when the alternative hypothesis is true.

Term Reports

Results for Factor B (Levels = 3)

				Multiply	SD of	Standard			
				Means	Effects	Deviation	Effect		
Test	Power	n	N	By	(Sm)	(Sigma)	Size	Alpha	Beta
GG F	0.3096	2	6	1.00	3.74	3.29	1.14	0.0500	0.6904
GG F	0.6459	3	9	1.00	3.74	3.29	1.14	0.0500	0.3541
GG F	0.8478	4	12	1.00	3.74	3.29	1.14	0.0500	0.1522
GG F	0.9415	5	15	1.00	3.74	3.29	1.14	0.0500	0.0585
GG F	0.9793	6	18	1.00	3.74	3.29	1.14	0.0500	0.0207
GG F	0.9931	7	21	1.00	3.74	3.29	1.14	0.0500	0.0069
GG F	0.9978	8	24	1.00	3.74	3.29	1.14	0.0500	0.0022

Results for Factor W (Levels = 4)

				Multiply	SD of	Standard			
				Means	Effects	Deviation	Effect		
Test	Power	n	N	By	(Sm)	(Sigma)	Size	Alpha	Beta
GG F	0.3266	2	6	1.00	1.92	1.31	1.47	0.0500	0.6734
GG F	0.8099	3	9	1.00	1.92	1.31	1.47	0.0500	0.1901
GG F	0.9486	4	12	1.00	1.92	1.31	1.47	0.0500	0.0514
GG F	0.9871	5	15	1.00	1.92	1.31	1.47	0.0500	0.0129
GG F	0.9970	6	18	1.00	1.92	1.31	1.47	0.0500	0.0030
GG F	0.9993	7	21	1.00	1.92	1.31	1.47	0.0500	0.0007
GG F	0.9999	8	24	1.00	1.92	1.31	1.47	0.0500	0.0001

Results for Term BW

				Multiply	SD of	Standard			
				Means	Effects	Deviation	Effect		
Test	Power	n	N	By	(Sm)	(Sigma)	Size	Alpha	Beta
GG F	0.1588	2	6	1.00	1.92	1.31	1.47	0.0500	0.8412
GG F	0.6267	3	9	1.00	1.92	1.31	1.47	0.0500	0.3733
GG F	0.8598	4	12	1.00	1.92	1.31	1.47	0.0500	0.1402
GG F	0.9535	5	15	1.00	1.92	1.31	1.47	0.0500	0.0465
GG F	0.9860	6	18	1.00	1.92	1.31	1.47	0.0500	0.0140
GG F	0.9961	7	21	1.00	1.92	1.31	1.47	0.0500	0.0039
GG F	0.9990	8	24	1.00	1.92	1.31	1.47	0.0500	0.0010

The *Term Reports* provide a complete report for each term at all sample sizes. They are especially useful when you are only interested in the power of one or two terms.

The definitions of each of the columns of the report are identical to the corresponding columns in the *Design Report*, so they are not repeated here.

Geisser-Greenhouse Correction Detail Report

			Critical					
Term (Levels)	Power	Alpha	F	Lambda	df1|df2	Epsilon	E(Epsilon)	G1
n = 2 N = 6 Means x 1								
B (3)	0.3096	0.0500	9.55	7.74	2|3	1.00	1.00	0.00
W (4)	0.3266	0.0500	7.60	12.88	3|9	0.77	0.43	-1.01
BW	0.1588	0.0500	6.92	12.88	6|9	0.77	0.43	-1.01
n = 3 N = 9 Means x 1								
B (3)	0.6459	0.0500	5.14	11.62	2|6	1.00	1.00	0.00
W (4)	0.8099	0.0500	4.12	19.32	3|18	0.77	0.60	-1.01
BW	0.6267	0.0500	3.46	19.32	6|18	0.77	0.60	-1.01
n = 4 N = 12 Means x 1								
B (3)	0.8478	0.0500	4.26	15.49	2|9	1.00	1.00	0.00
W (4)	0.9486	0.0500	3.58	25.76	3|27	0.77	0.66	-1.01
BW	0.8598	0.0500	2.95	25.76	6|27	0.77	0.66	-1.01
n = 5 N = 15 Means x 1								
B (3)	0.9415	0.0500	3.89	19.36	2|12	1.00	1.00	0.00
W (4)	0.9871	0.0500	3.36	32.20	3|36	0.77	0.68	-1.01
BW	0.9535	0.0500	2.75	32.20	6|36	0.77	0.68	-1.01

```
n = 6  N = 18 Means x 1
B (3)           0.9793    0.0500    3.68     23.23     2|15     1.00      1.00      0.00
W (4)           0.9970    0.0500    3.25     38.64     3|45     0.77      0.70     -1.01
BW              0.9860    0.0500    2.64     38.64     6|45     0.77      0.70     -1.01
n = 7  N = 21 Means x 1
B (3)           0.9931    0.0500    3.55     27.11     2|18     1.00      1.00      0.00
W (4)           0.9993    0.0500    3.17     45.07     3|54     0.77      0.71     -1.01
BW              0.9961    0.0500    2.57     45.07     6|54     0.77      0.71     -1.01
n = 8  N = 24 Means x 1
B (3)           0.9978    0.0500    3.47     30.98     2|21     1.00      1.00      0.00
W (4)           0.9999    0.0500    3.12     51.51     3|63     0.77      0.72     -1.01
BW              0.9990    0.0500    2.52     51.51     6|63     0.77      0.72     -1.01
```

This report gives the details of the components of the Geisser-Greenhouse correction for each term and sample size. It is useful when you want to compare various aspects of this test.

The definitions of each of the columns of the report are as follows.

Term

This column contains the identifying label of the term. For factors, the number of levels is also given in parentheses.

Power

This is the computed power for the term.

Alpha

Alpha is the significance level of the test.

Critical F

This is the critical value of the F statistic. An *F* value computed from the data that is larger than this value is statistically significant at the alpha level given.

Lambda

This is the value of the noncentrality parameter λ of the approximate noncentral *F* distribution.

df1|df2

These are the values of the numerator and denominator degrees of freedom of the approximate *F* test that is used. These values are useful when comparing various designs. Other things being equal, you would like to have df2 large and df1 small.

Epsilon

The Geisser-Greenhouse epsilon is a measure of how far the covariance matrix departs from the assumption of circularity.

E(Epsilon)

This is the expected value of epsilon. It is a measure of how far the covariance matrix departs from the assumption of circularity.

G1

G1 is part of a correction factor used to convert ε to $E\left(\hat{\varepsilon}\right)$. It is reported for your convenience.

Summary Statements

```
A repeated measures design with 1 between factor and 1 within factor has 3 groups with 2
subjects each for a total of 6 subjects. Each subject is measured 4 times. This design achieves
31% power to test factor B if a Geisser-Greenhouse Corrected F Test is used with a 5%
significance level and the actual effect standard deviation is 3.74 (an effect size of 1.14),
achieves 33% power to test factor W if a Geisser-Greenhouse Corrected F Test is used with a 5%
significance level and the actual effect standard deviation is 1.92 (an effect size of 1.47),
and achieves 16% power to test the BW interaction if a Geisser-Greenhouse Corrected F Test is
used with a 5% significance level and the actual effect standard deviation is 1.92 (an effect
size of 1.47).
```

A summary statement can be generated for each sample size that is used. This statement gives the results in sentence form. The number of designs reported on textually is controlled by the Summary Statement option on the Reports Tab.

Means Matrix

```
Name           B1          B2          B3
W1         -10.62        5.19       -2.72
W2           1.46        3.97        2.72
W3          -4.58        4.58        0.00
W4          -4.58        4.58        0.00
```

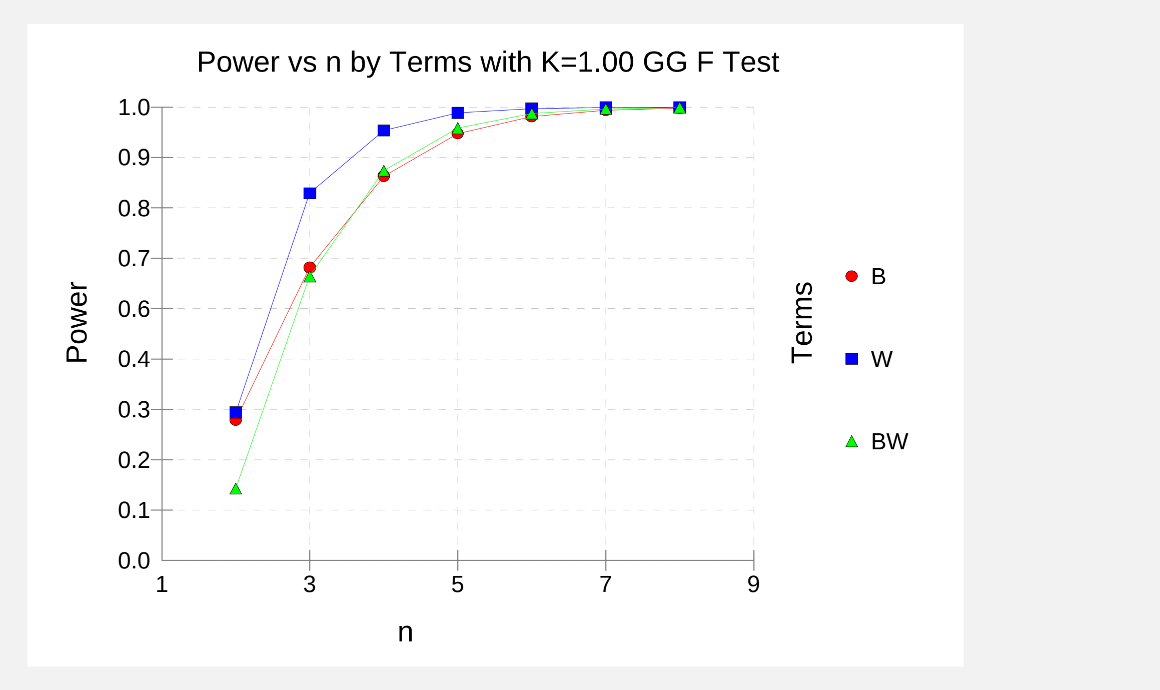
This report shows the means matrix that was read in from the spreadsheet or generated by the Sm values that were given. It may be used to get an impression of the magnitude of the difference among the means that is being studied. When a Means Multiplier, *K*, is used, each value of *K* is multiplied times each value of this matrix.

Variance-Covariance Matrix Section

```
Variance-Covariance Matrix Section
Time           W1          W2          W3           W4
W1           4.00        0.70        0.49         0.34
W2           0.70        4.00        0.70         0.49
W3           0.49        0.70        4.00         0.70
W4           0.34        0.49        0.70         4.00
```

This report shows the variance-covariance matrix that was read in from the spreadsheet or generated by the settings of on the Covariance tab. The standard deviations are given on the diagonal and the autocorrelations are given off the diagonal.

Plots Section



The chart shows the relationship between power and *n* for the terms in the design. Note that high-order interactions may be omitted from the plot by reducing the Max Term-Order Plotted option on the Plot Setup tab.

Example 2 – Varying the Difference Between the Means

Continuing with Example 1, the researchers want to evaluate the impact on power of varying the size of the difference among the means for a range of sample sizes from 2 to 8 per groups. The researchers could try calculating various multiples of the means, inputting them, and recording the results. This can be accomplished very easily by using the *K* option.

Keeping all other settings as in Example 1, the value of *K* is varied from 0.2 to 3.0 in steps of 0.2. We determined these values by experimentation so that a full range of power values are shown on the plots.

In the output to follow, we only display the plots. You may want to display the numeric reports as well, but we do not here in order to save space.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	**Value**
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**2 3 4 8**
=n's	**checked**
Means Matrix	blank
K (Means Multipliers)	**0.2 to 3.0 by 0.2**
For First Between-Subject Factor	
Label	**B**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**93 87 84**
For First Within-Subject Factor	
Label	**W**
Levels	**4**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**93 89 88 91**
Interactions Tab	
2-Way(Mixed) Interaction	
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**W 100**
Covariance Tab	
Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**4**
Time Metric	*Ignored*
Specify How Autocorr's Change	**1st Order Autocorr**
R1 (Autocorrelation)	**0.7**
Max Time Diff.	**100** (This large value will cause R2 to be ignored.)
Reports Tab	
Numeric Results by Term	**Not checked**
Numeric Results by Design	**Not checked**
Regular F Test	**Not checked**
GG F Test	**Not checked**
Wilks' Lambda	**Not checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
GG Detail Report	**Not checked**
Means Matrix	**Not checked**
Covariance Matrix	**Not checked**

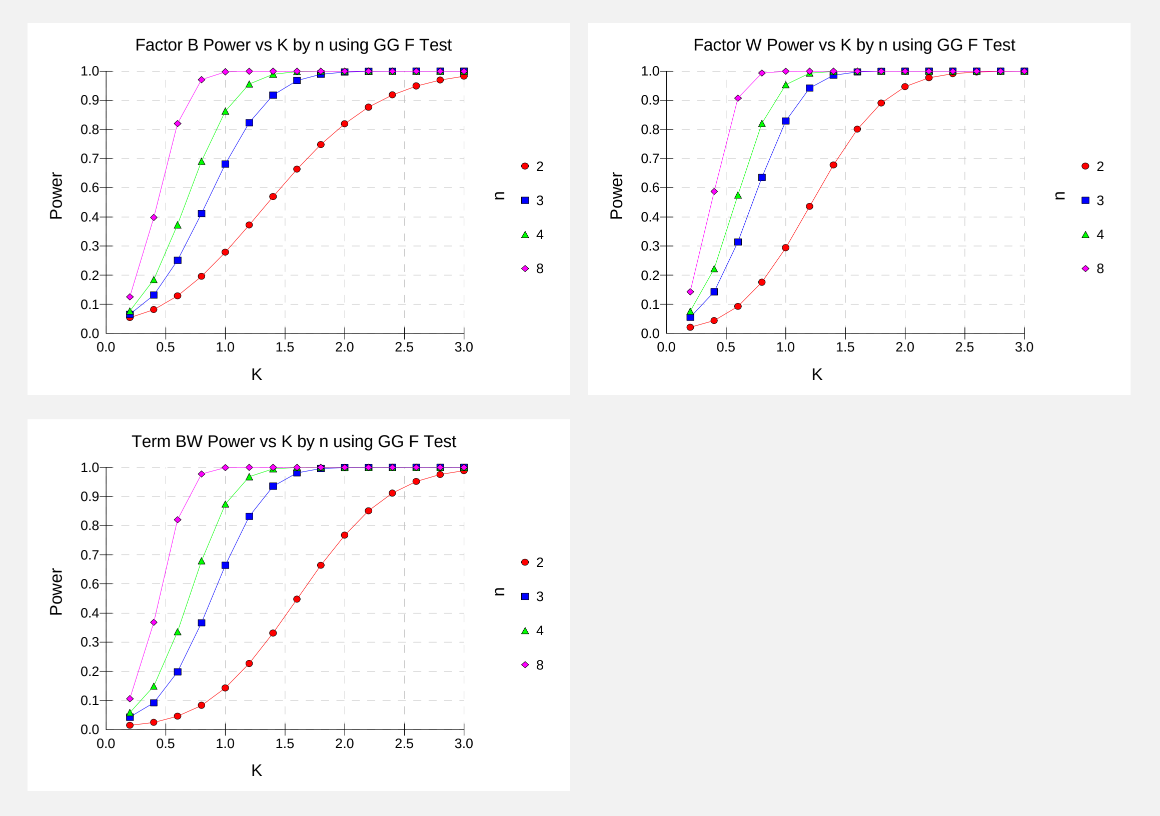
Reports Tab (continued)

Show Plot 1 .... **Not checked**
Show Plot 2 .... **Checked**
Test That is Plotted .... **GG F Test**
Max Term-Order Plotted **2**

Output

Click the Run button to perform the calculations and generate the following output.

Plots Section



These charts show how the power depends on the relative size of *Sm* (i.e. *K*) as well as the group sample size *n*.

Example 3 – Impact of the Number of Repeated Measurements

Continuing with Example 1, the researchers want to study the impact on power of changing the number of measurements made on each individual. Their experimental protocol calls for four measurements that are 30 minutes apart. They want to see the impact of taking twice that many measurements. To keep the output simple and two the point, they decide to look at the case when *n* = 4 and *K* = 1.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**4**
=n's	**checked**
Means Matrix	blank
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**B**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**93 87 84**
For First Within-Subject Factor	
Label	**W**
Levels	**4**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**RANGE 88 93**
Interactions Tab	
2-Way(Mixed) Interaction	
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**W 100**
Covariance Tab	
Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**4**
Time Metric	*Ignored*
Specify How Autocorr's Change	**1st Order Autocorr**

Covariance Tab (continued)

R1 (Autocorrelation)................................**0.7**
Max Time Diff. ....**100** (This large value will cause R2 to be ignored.)

Reports Tab

Numeric Results by Term.......................**Not checked**
Numeric Results by Design....................**Checked**
Regular F Test ....**Not checked**
GG F Test..**Checked**
Wilks' Lambda..**Not checked**
Pillai-Bartlett ....**Not checked**
Hotelling-Lawley.....................................**Not checked**
GG Detail Report....................................**Not checked**
Means Matrix..**Not checked**
Covariance Matrix ....**Not checked**
Show Plot 1 ....**Not checked**
Show Plot 2 ....**Not checked**
Max Term-Order Reported......................**2**

Output

Click the Run button to perform the calculations and generate the following output.

Design Report

```
                                         Multiply  SD of  Standard
                                           Means Effects Deviation  Effect
Term    Test    Power     n      N          By    (Sm)   (Sigma)    Size     Alpha     Beta
(Results with 4 measurements)
B(3)    GG F    0.8478    4     12        1.00    3.74      3.29    1.14    0.0500   0.1522
W(4)    GG F    0.9350    4     12        1.00    1.86      1.31    1.42    0.0500   0.0650
BW      GG F    0.8348    4     12        1.00    1.86      1.31    1.42    0.0500   0.1652

(Results with 8 measurements)
B(3)    GG F    0.8375    4     12        1.00    3.74      3.34    1.12    0.0500   0.1625
W(8)    GG F    0.9354    4     12        1.00    1.64      0.83    1.97    0.0500   0.0646
BW      GG F    0.8321    4     12        1.00    1.64      0.83    1.97    0.0500   0.1679
```

Notice that the power of the between subjects factor decreased slightly, the power of the within-subjects factor increased slightly, and the power of the interaction test decreased slightly. This pattern of increase or decrease depends on all the settings.

We tried varying the value of the autocorrelation from 0.7 to 0.1 and found this to impact the direction of the change in the number of measurements. Hence, our conclusion is that there is no single answer. Changing the number of measurements may increase or decrease the power of a specific test depending on the values of the other parameters.

Example 4 – Power after a Study

This example will show how to calculate the power of *F* tests from data that have already been collected and analyzed using the analysis of variance. The following results were obtained using the analysis of variance procedure in ***NCSS***. In this example, Gender is the between factor with two levels and Treatment is the within factor with three levels. The experiment was conducted with two subjects per group, but there is interest in the power for 2, 3, and 4 subjects per group. All significance levels are set to 0.05.

```
Analysis of Variance Table
Source                         Sum of          Mean                        Prob
Term                DF         Squares         Square        F-Ratio       Level
A (Gender)           1         21.33333        21.33333      32.00         0.029857
B(A)                 2         1.333333        0.6666667
C (Treatment)        2         5.166667        2.583333      6.20          0.059488
AC                   2         5.166667        2.583333      6.20          0.059488
BC(A)                4         1.666667        0.4166667
Total (Adjusted)    11         34.66667
Total               12

Means and Effects Section
                                           Standard
Term           Count        Mean           Error
All               12    17.33333
A: Gender
Females            6          16           0.3333333
Males              6    18.66667           0.3333333
C: Treatment
L                  4       16.75           0.3227486
M                  4          17           0.3227486
H                  4       18.25           0.3227486
AC: Gender,Treatment
Females,L          2        14.5           0.4564355
Females,M          2          16           0.4564355
Females,H          2        17.5           0.4564355
Males,L            2          19           0.4564355
Males,M            2          18           0.4564355
Males,H            2          19           0.4564355
```

Note that the treatment means (L, M, and H) show an increasing pattern from 16.75 to 18.25, but the hypothesis test of this factor is not statistically significant at the 0.05 level. We will now calculate the power of the three *F* tests using ***PASS***. We will use the regular F test since that is what was used in the above table.

From the above printout, we note that MSB = 0.6666667 and MSW = 0.4166667. Plugging these values into the estimating equations

$$\hat{\rho} = \frac{MSB - MSW}{MSB + \left(T-1\right)MSW}$$

$$\hat{\sigma}^2 = \frac{MSW}{1-\hat{\rho}}$$

yields

$$\hat{\rho} = \frac{0.6666667 - 0.4166667}{0.6666667 + (3-1)0.4166667} = 0.16666667$$

$$\hat{\sigma}^2 = \frac{0.4166667}{1 - 0.16666667} = 0.5$$

so that

$$\hat{\sigma} = \sqrt{0.5} = 0.70710681$$

With these values calculated, we can setup ***PASS*** to calculate the power of the three *F* tests as follows.

Setup

This section presents the values of each of the parameters needed to run this example. First, you will need to open the PASS RM Example 2.S0 dataset by clicking on the **DATA** icon in the toolbar at the top of PASS Home window. On the Data window, select **File**, then **Open** from the menus. Navigate to the **Data** folder that is located in the folder into which you installed ***PASS*** and select **PASS RM Example 2.S0**. Once the dataset is loaded, go back to the PASS Home window. From the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**2 3 4**
=n's	**checked**
Means Matrix	**FEMALES-MALES**
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**B**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	*Ignored since the Means Matrix is loaded.*
For First Within-Subject Factor	
Label	**W**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	*Ignored since the Means Matrix is loaded.*
Interactions Tab	
2-Way(Mixed) Interaction	
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	*Ignored since the Means Matrix is loaded.*

Covariance Tab

Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**.70710681**
Time Metric	*Ignored*
Specify How Autocorr's Change	**Constant**
R1 (Autocorrelation)	**0.16666667**
Max Time Diff.	**100** (This large value will cause R2 to be ignored.)

Reports Tab

Numeric Results by Term	**Checked**
Numeric Results by Design	**Not Checked**
Regular F Test	**Checked**
GG F Test	**Not checked**
Wilks' Lambda	**Not checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
GG Detail Report	**Not checked**
Means Matrix	**Not checked**
Covariance Matrix	**Not checked**
Test in Summary Statement	**Regular F Test**
Show Plot 1	**Not checked**
Show Plot 2	**Not checked**
Max Term-Order Reported	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                                             Multiply    SD of  Standard
                                                Means  Effects  Deviation   Effect
Term      Test      Power      n       N          By      (Sm)    (Sigma)     Size      Alpha        Beta
B(2)         F     0.8004      2       4        1.00      1.33       0.47     2.83     0.0500      0.1996
W(3)         F     0.5536      2       4        1.00      0.66       0.37     1.76     0.0500      0.4464
BW           F     0.5536      2       4        1.00      0.66       0.37     1.76     0.0500      0.4464

B(2)         F     0.9985      3       6        1.00      1.33       0.47     2.83     0.0500      0.0015
W(3)         F     0.8933      3       6        1.00      0.66       0.37     1.76     0.0500      0.1067
BW           F     0.8933      3       6        1.00      0.66       0.37     1.76     0.0500      0.1067

B(2)         F     1.0000      4       8        1.00      1.33       0.47     2.83     0.0500      0.0000
W(3)         F     0.9801      4       8        1.00      0.66       0.37     1.76     0.0500      0.0199
BW           F     0.9801      4       8        1.00      0.66       0.37     1.76     0.0500      0.0199
```

You can see that the power of the tests on W and BW was only 0.55 for an *n* of 2. However, if *n* would have been 3, a much more reasonable power of 0.89 would have been achieved.

Example 5 – Cross-over Design

A *crossover design* is a special type of repeated measures design in which the treatments are applied to the subjects in different orders. The between-subjects (grouping) factor is defined by the specific sequence in which the treatments are applied. For example, suppose the treatments are represented by B1 and B2. Further suppose that half the subjects receive treatment B1 followed by treatment B2 (sequence B1B2), while the other half receive treatment B2 followed by treatment B1 (sequence B2B1). This is a two-group crossover design.

Crossover designs assume that a long enough period elapses between measurements so that the effects of one treatment are *washed out* before the next treatment is applied. This is known as the assumption of no *carryover* effects.

When a crossover design is analyzed using repeated measures, the interaction is the only term of interest. The *F* test on the between factor tests whether averages across each sequence are equal—a test of little interest. The *F* test on the within factor tests whether the response is different across the time periods—also of little interest. The *F* test for interaction tests whether the change in response across time is the same for both sequences. The interaction can only be significant if the treatments effect the outcome differently. Hence, to specify a crossover design requires the careful specification of the interaction effects.

With this background, we present an example. Suppose researchers want to investigate the reduction in heart-beat rate caused by the administration of a certain drug using a simple two-period crossover design. The researchers want a sample size large enough to detect a drop in heart-beat rate from 95 to 90 with a power of 90% at the 0.05 significance level. Previous studies have shown a within-patient autocorrelation of 0.50 and a standard deviation of 3.98. They decide to consider sample sizes between 2 and 8.

The hypothesized interaction is specified by entering the mean heart-beat rates of the four treatment groups as 95, 90, 90, and 95. Since the standard deviation of these values is all that is used, the order of these values does not matter. In this case the sequences means are both 92.5 and the average time-period means are both 92.5. Hence, the interaction effects are 2.5, -2.5, -2.5, and 2.5. You can check that the set of numbers '95, 90, 90, 95' has the same standard deviation as the set '2.5, -2.5, -2.5, 2.5' or even '5, 0, 0, 5'. All of these will work.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**2 to 8 by 1**
=n's	**checked**
Means Matrix	blank
K (Means Multipliers)	**1.0**

Data Tab (continued)

For First Between-Subject Factor
Label..**Seq**
Levels..**2**
Alpha **0.05**
Power ....*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects).........**0.1** (This value is arbitrary.)

For First Within-Subject Factor
Label..**Time**
Levels..**2**
Alpha **0.05**
Power ....*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects).........**0.1** (This value is arbitrary.)

Interactions Tab

2-Way(Mixed) Interaction
Alpha **0.05**
Power ....*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects).........**95 90 90 95**

Covariance Tab

Specify Covariance Method ....**1) Standard Deviations and Autocorrelations**
How SD’s Change Across Time..............**Constant**
SD1 (Standard Deviation 1) **3.98**
Time Metric ....*Ignored*
Specify How Autocorr’s Change ....**1st Order Autocorr**
R1 (Autocorrelation)................................**0.5**
Max Time Diff. ....**100** (This large value will cause R2 to be ignored.)

Reports Tab

Numeric Results by Term........................**Checked**
Numeric Results by Design.....................**Not checked**
Regular F Test ....**Not checked**
GG F Test ....**Checked**
Wilks’ Lambda...**Not checked**
Pillai-Bartlett ....**Not checked**
Hotelling-Lawley......................................**Not checked**
GG Detail Report.....................................**Not checked**
Means Matrix...**Not checked**
Covariance Matrix ....**Not checked**
Show Plot 1 ....**Not checked**
Show Plot 2 ....**Not checked**
Max Term-Order Reported......................**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Results for Term SeqTime
                                    Multiply      SD of    Standard
                                       Means    Effects   Deviation    Effect
Test      Power     n        N            By       (Sm)     (Sigma)      Size       Alpha       Beta
GG F      0.3017    2        4          1.00       2.50        1.99      1.26      0.0500     0.6983
GG F      0.6401    3        6          1.00       2.50        1.99      1.26      0.0500     0.3599
GG F      0.8395    4        8          1.00       2.50        1.99      1.26      0.0500     0.1605
GG F      0.9338    5       10          1.00       2.50        1.99      1.26      0.0500     0.0662
GG F      0.9742    6       12          1.00       2.50        1.99      1.26      0.0500     0.0258
GG F      0.9903    7       14          1.00       2.50        1.99      1.26      0.0500     0.0097
GG F      0.9965    8       16          1.00       2.50        1.99      1.26      0.0500     0.0035
```

We only display the interaction term since that is the only term of interest. A quick glance at the plot shows that 90% power is achieved when *n* is five. This corresponds to a total sample size of ten subjects.

Example 6 – Power of a Completed Cross-over Design

The following analysis of variance table was generated by ***NCSS*** for a set of crossover data. Find the power of the interaction *F* test assuming a significance level of 0.05.

```
Analysis of Variance Table
Source                               Sum of          Mean                     Prob
Term                      DF        Squares        Square    F-Ratio        Level
A: Sequence                1        89397.6       89397.6       1.19     0.285442
B(A): Subject             28        2110739      75383.54
C: Period                  1       117395.3      117395.3       1.40     0.246854
AC                         1       122401.7      122401.7       1.46     0.237263
BC(A)                     28        2349752      83919.72
Total (Adjusted)          59        4789686
Total                     60

Means Section
                                                           Standard
Term                   Count            Mean                  Error
All                       60        492.2000
A: Sequence
1                         30        453.6000               50.12768
2                         30        530.8000               50.12768
C: Period
1                         30        447.9667               52.88973
2                         30        536.4333               52.88973
AC: Sequence,Period
1,1                       15        364.2000               74.79738
1,2                       15        543.0000               74.79738
2,1                       15        531.7333               74.79738
2,2                       15        529.8666               74.79738
```

One difficulty in analyzing an existing crossover design is determining an appropriate value for the hypothesized interaction effects. One method is to find the standard deviation of the interaction effects by taking the square root of the Sum of Squares for the interaction divided by the total number of observations. In this case,

$$\begin{aligned}\sigma_{Interaction} &= \sqrt{\frac{122401.7}{60}} \\ &= 45.1667\end{aligned}$$

Another method is to find the individual interaction effects by subtraction. This method proceeds as follows.

First, subtract the Period means from the Sequence by Period means.

$$\begin{bmatrix} 364.2000 & 531.7333 \\ 543.0000 & 529.8666 \end{bmatrix} - \begin{bmatrix} 447.9667 \\ 536.4333 \end{bmatrix} = \begin{bmatrix} -83.7667 & 83.7667 \\ 6.5667 & -6.5667 \end{bmatrix}$$

Next, compute the column means and subtract them from the current values. This results in the effects.

$$\begin{bmatrix} -83.7667 & 83.7667 \\ 6.5667 & -6.5667 \end{bmatrix} - \begin{bmatrix} -38.6000 & 38.6000 \\ -38.6000 & 38.6000 \end{bmatrix} = \begin{bmatrix} -45.1667 & 45.1667 \\ 45.1667 & -45.1667 \end{bmatrix}$$

Finally, compute the standard deviation of the effects. Since the mean of the effects is zero, the standard deviation is

$$\begin{aligned}\sigma_{Interaction} &= \sqrt{\frac{\left(-45.1667\right)^2 + \left(45.1667\right)^2 + \left(45.1667\right)^2 + \left(-45.1667\right)^2}{4}} \\ &= 45.1667\end{aligned}$$

Another difficulty that must be solved is to estimate the autocorrelation and within-subject standard deviation. From the above printout, we note that MSB = 75383.54 and MSW = 83919.72. Plugging these values into the estimating equations

$$\hat{\rho} = \frac{MSB - MSW}{MSB + \left(T-1\right)MSW}$$

$$\hat{\sigma}^2 = \frac{MSW}{1-\hat{\rho}}$$

yields

$$\hat{\rho} = \frac{75383.54 - 83919.72}{75383.54 + (2-1)83919.72} = -0.05358447$$

$$\hat{\sigma}^2 = \frac{83919.72}{1 + 0.05358447} = 79651.63$$

so that

$$\hat{\sigma} = \sqrt{79651.63} = 282.2262$$

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**15**
=n's	**checked**
Means Matrix	blank
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**S**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**453.6000 530.8000**
For First Within-Subject Factor	
Label	**P**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**447.9667 536.4333**
Interactions Tab	
2-Way(Mixed) Interaction	
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	**45.1667**
Covariance Tab	
Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**282.2262**
Time Metric	*Ignored*
Specify How Autocorr's Change	**1st Order Autocorr**
R1 (Autocorrelation)	**-0.05358447**
Max Time Diff.	**100** (This large value will cause R2 to be ignored.)
Reports Tab	
Numeric Results by Term	**Not checked**
Numeric Results by Design	**Checked**
Regular F Test	**Not checked**
GG F Test	**Checked**
Wilks' Lambda	**Not checked**
Pillai-Bartlett	**Not checked**

Reports Tab (continued)

Hotelling-Lawley	**Not checked**
GG Detail Report	**Not checked**
Means Matrix	**Not checked**
Covariance Matrix	**Not checked**
Number of Summary Statements	**0**
Show Plot 1	**Not checked**
Show Plot 2	**Not checked**
Max Term-Order Reported	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                                            Multiply   SD of  Standard
                                               Means Effects Deviation    Effect
Term     Test     Power       n       N          By    (Sm)   (Sigma)      Size       Alpha        Beta
S(2)     GG F    0.1832      15      30        1.00   38.60    194.14      0.20      0.0500      0.8168
P(2)     GG F    0.2078      15      30        1.00   44.23    204.84      0.22      0.0500      0.7922
SP       GG F    0.2147      15      30        1.00   45.17    204.84      0.22      0.0500      0.7853
```

Notice that these power values are low. Fifteen was not a large enough sample size to detect Sm values near 40.

Example 7 – Validation using O'Brien and Muller

O'Brien and Muller's article in the book edited by Edwards (1993) analyze the power of a two-group repeated-measures experiment in which three measurements are made on each subject.

The hypothesized means are

	Group 1	Group 2
Time 1	3	1
Time 2	12	5
Time 3	8	7

The covariance matrix is

	Time 1	Time 2	Time 3
Time 1	25	16	12
Time 2	16	64	30
Time 3	12	30	36

With *n*'s of 12, 18, and 24 and an alpha of 0.05, they obtained power values using the Wilks' Lambda test. Their reported power values are

Power Values for each Term

n	Group	Time	Interaction
12	0.326	0.983	0.461
18	0.467	0.999	0.671
24	0.589	0.999	0.814

O'Brien, in a private communication, re-ran these data using the Geisser-Greenhouse correction. His results were as follows:

Power Values for each Term

n	Group	Time	Interaction
12	0.326	0.993	0.486
18	0.467	0.999	0.685
24	0.589	0.999	0.819

In order to run this example in ***PASS***, the values of the means and the covariance matrix (given above) must be entered on a spreadsheet. We have loaded these values into the database called OBRIEN. Either enter the values yourself, or load the OBRIEN database which should be in the Data directory. The instructions below assume that the means are in columns one and two, while the covariance matrix is in columns four through six of the current database.

Setup

In order to run this example the **Obrien.S0** data must be loaded into the spreadsheet. To open the spreadsheet window from the PASS Home Window, click on the **Tools** menu and select **Spreadsheet**. Once the spreadsheet is open, the **Obrien.S0** data is loaded by clicking the **File** menu and selecting **Open**. The **Obrien.S0** file is then selected from the **DATA** folder (the default location for this folder is *C:\...\[My] Documents\NCSS\PASS2008*). Then click **Open**.

This section presents the values of each of the parameters needed to run this example. From the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example7** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**12 18 24**
=n's	**checked**
Means Matrix	**M1-M2**
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**G**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	*Ignored*

Data Tab (continued)

For First Within-Subject Factor

Label	**T**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	*Ignored*

Interactions Tab

2-Way(Mixed) Interaction

Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Sm (Standard Deviation of Effects)	*Ignored*

Covariance Tab

Specify Covariance Method	**2) Covariance Matrix Variables**
Spreadsheet Columns	**S1-S3**

Reports Tab

Numeric Results by Term	**Checked**
Numeric Results by Design	**Not checked**
Regular F Test	**Not checked**
GG F Test	**Checked**
Wilks' Lambda	**Checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
GG Detail Report	**Not checked**
Means Matrix	**Not checked**
Covariance Matrix	**Not checked**
Number of Summary Statements	**0**
Show Plot 1	**Not checked**
Show Plot 2	**Not checked**
Max Term-Order Reported	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Results for Factor G (Levels =2)
                                  Multiply    SD of    Standard
                                     Means  Effects   Deviation   Effect
Test       Power    n      N           By     (Sm)     (Sigma)     Size       Alpha      Beta
GG F      0.3263   12     24         1.00     1.67        5.17     0.32      0.0500    0.6737
Wilks     0.3263   12     24         1.00     1.67        5.17     0.32      0.0500    0.6737
GG F      0.4673   18     36         1.00     1.67        5.17     0.32      0.0500    0.5327
Wilks     0.4673   18     36         1.00     1.67        5.17     0.32      0.0500    0.5327
GG F      0.5889   24     48         1.00     1.67        5.17     0.32      0.0500    0.4111
Wilks     0.5889   24     48         1.00     1.67        5.17     0.32      0.0500    0.4111
```

```
Results for Factor T (Levels =3)
                                     Multiply      SD of   Standard
                                        Means    Effects  Deviation   Effect
Test       Power     n         N          By       (Sm)    (Sigma)     Size      Alpha        Beta
GG F      0.9933    12        24        1.00       2.86       2.73     1.05     0.0500      0.0067
Wilks     0.9825    12        24        1.00       2.86       2.73     1.05     0.0500      0.0175
GG F      0.9999    18        36        1.00       2.86       2.73     1.05     0.0500      0.0001
Wilks     0.9995    18        36        1.00       2.86       2.73     1.05     0.0500      0.0005
GG F      1.0000    24        48        1.00       2.86       2.73     1.05     0.0500      0.0000
Wilks     1.0000    24        48        1.00       2.86       2.73     1.05     0.0500      0.0000

Results for Term GT
                                     Multiply      SD of   Standard
                                        Means    Effects  Deviation   Effect
Test       Power     n         N          By       (Sm)    (Sigma)     Size      Alpha        Beta
GG F      0.4861    12        24        1.00       1.31       2.73     0.48     0.0500      0.5139
Wilks     0.4605    12        24        1.00       1.31       2.73     0.48     0.0500      0.5395
GG F      0.6850    18        36        1.00       1.31       2.73     0.48     0.0500      0.3150
Wilks     0.6706    18        36        1.00       1.31       2.73     0.48     0.0500      0.3294
GG F      0.8193    24        48        1.00       1.31       2.73     0.48     0.0500      0.1807
Wilks     0.8136    24        48        1.00       1.31       2.73     0.48     0.0500      0.1864
```

PASS agrees exactly with O’Brien’s calculations.

Example 8 – Unequal Group Sizes

Usually, in the planning stages, the group sample sizes are equal. Occasionally, however, you may want to plan for a situation in which one group will have a much larger sample size than the others. Also, when doing a power analysis on a study that has already been conducted, the group sample sizes are often unequal.

In this example, we will re-analyze the Example 4. However, we will now assume that there were four subjects in group 1 and eight subjects in group 2. The setup and output for this example are as follows. Remember that you must open the database PASS RM EXAMPLE2 before running this example.

Setup

In order to run this example the **PASS RM Example 2.S0** data must be loaded into the spreadsheet. To open the spreadsheet window from the PASS Home Window, click on the **Tools** menu and select **Spreadsheet**. Once the spreadsheet is open, the **PASS RM Example 2.S0** data is loaded by clicking the **File** menu and selecting **Open**. The **PASS RM Example 2.S0** file is then selected from the **DATA** folder (the default location for this folder is *C:\...\[My] Documents\NCSS\PASS2008*). Then click **Open**.

This section presents the values of each of the parameters needed to run this example. From the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example8** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**4 8**
=n's	**Not checked**
Means Matrix	**FEMALES-MALES**
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**B**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	*Ignored since the Means Matrix is loaded.*
For First Within-Subject Factor	
Label	**W**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	*Ignored since the Means Matrix is loaded.*
Interactions Tab	
2-Way(Mixed) Interaction	
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	*Ignored since the Means Matrix is loaded.*
Covariance Tab	
Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**.70710681**
Time Metric	*Ignored*
Specify How Autocorr's Change	**Constant**
R1 (Autocorrelation)	**0.16666667**
Max Time Diff.	**100** (This large value will cause R2 to be ignored.)
Reports Tab	
Numeric Results by Term	**Checked**
Numeric Results by Design	**Not Checked**
Regular F Test	**Not checked**
GG F Test	**Checked**
Wilks' Lambda	**Not checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
GG Detail Report	**Not checked**
Means Matrix	**Not checked**
Covariance Matrix	**Not checked**
Show Plot 1	**Not checked**
Show Plot 2	**Not checked**
Max Term-Order Reported	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
                                               Multiply   SD of  Standard
                                                  Means Effects Deviation  Effect
Term     Test       Power     n       N             By     (Sm)   (Sigma)    Size     Alpha       Beta
B(2)     GG F      1.0000   6.0      12           1.00     1.26      0.47    2.67    0.0500     0.0000
W(3)     GG F      0.9983   6.0      12           1.00     0.62      0.37    1.66    0.0500     0.0017
BW       GG F      0.9983   6.0      12           1.00     0.62      0.37    1.66    0.0500     0.0017
n's: 4 8
```

Notice that the values of *n* are now shown to one decimal place. That is because the value reported is the average value of *n*. The actual *n*'s are shown following the report.

Example 9 – Designs with More Than Two Factors

Occasionally, you will have a design that has more than two factors. We will now show you how to compute the necessary sample size for such a design.

Suppose your design calls for two between-subject factors, Age (A) and Gender (G), and two within-subject factors, Dose-Level (D) and Application-Method (M). Suppose the number of levels of these four factors are, respectively, 3, 2, 4, and 2.

Our first task is to determine appropriate values of Sm for each of the terms. We decide to ignore the interactions during the planning and only consider the factors themselves. The desired difference to be detected among the three age groups can be represented by the means 80, 88, and 96. The desired difference to be detected among the two genders can be represented by the means 80 and 96. The desired difference to be detected among the four dose levels is represented by the means 80, 82, 84, and 86. The desired difference to be detected among the two application methods is represented by the means 80 and 86.

Our next task is to specify the covariance matrix. From previous experience, we have found that a constant value of 20.0 is appropriate for SD1. An autocorrelation of 0.5 with a first-order autocorrelation pattern is also appropriate.

Finally, we decide to calculate the power using the GG F test at the following sample sizes: 2, 4, 6, 8, 10, 20, 30, and 40.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Repeated Measures Analysis of Variance** procedure window by clicking on **Means**, then **Many Means (ANOVA)**, then **Repeated Measures ANOVA**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example9** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**2 4 6 8 10 20 30 40**
=n's	**Checked**
Means Matrix	*blank*
K (Means Multipliers)	**1.0**
For First Between-Subject Factor	
Label	**A**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	**80 88 96**
For Second Between-Subject Factor	
Label	**G**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	**80 96**
For First Within-Subject Factor	
Label	**D**
Levels	**4**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	**80 82 84 86**
For Second Within-Subject Factor	
Label	**M**
Levels	**2**
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	**80 86**
Interactions Tab	
All Interactions	
Alpha	**0.05**
Power	*Ignored since this is the Find setting.*
Sm (Standard Deviation of Effects)	**D 100**
Covariance Tab	
Specify Covariance Method	**1) Standard Deviations and Autocorrelations**
How SD's Change Across Time	**Constant**
SD1 (Standard Deviation 1)	**20**
Time Metric	*Ignored*
Specify How Autocorr's Change	**1st Order Autocorrelation**
R1 (Autocorrelation)	**0.5**
Max Time Diff.	**100** (This large value will cause R2 to be ignored.)

Reports Tab

Numeric Results by Term........................**Not checked**
Numeric Results by Design....................**Checked**
Regular F Test ....**Not checked**
GG F Test...**Checked**
Wilks' Lambda..**Not checked**
Pillai-Bartlett ....**Not checked**
Hotelling-Lawley.....................................**Not checked**
GG Detail Report....................................**Not checked**
Means Matrix..**Not checked**
Covariance Matrix ....**Not checked**
Test in Summary Statement....................**Regular F Test**
Show Plot 1 ....**Checked**
Show Plot 2 ....**Not checked**
Test That is Plotted ....**GG F Test**
Max Term-Order Plotted **1**
Max Term-Order Reported......................**1**

Output

Click the Run button to perform the calculations and generate the following output.

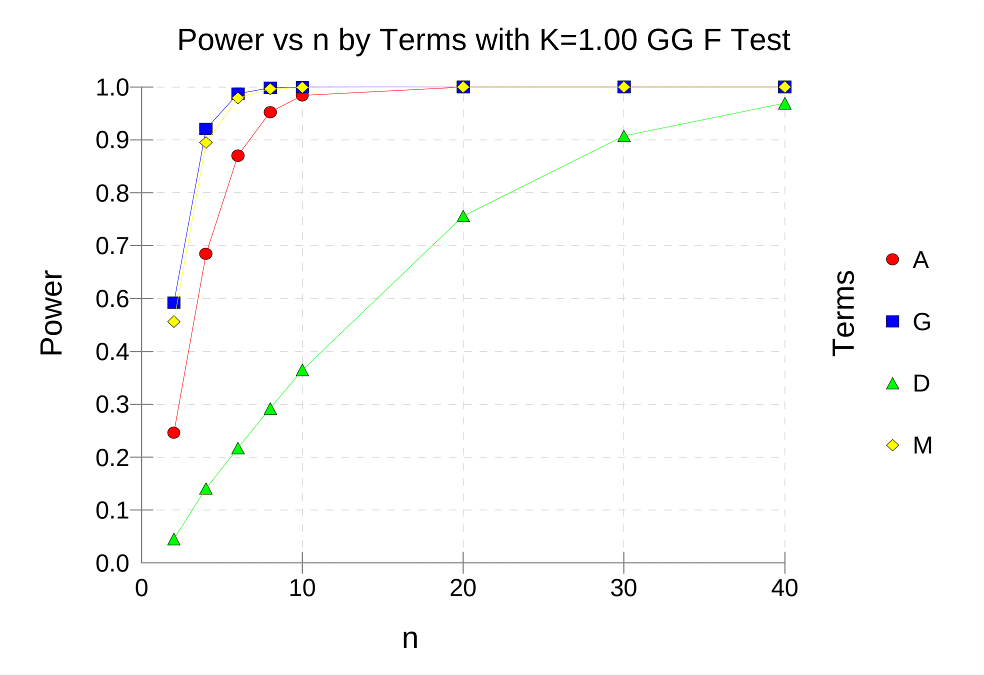
Numeric Results

					Multiply	SD of	Standard			
					Means	Effects	Deviation	Effect		
Term	Test	Power	n	N	By	(Sm)	(Sigma)	Size	Alpha	Beta
A(3)	GG F	0.2735	2	12	1.00	6.53	11.18	0.58	0.0500	0.7265
G(2)	GG F	0.5465	2	12	1.00	8.00	11.18	0.72	0.0500	0.4535
D(4)	GG F	0.0506	2	12	1.00	2.24	7.64	0.29	0.0500	0.9494
M(2)	GG F	0.5072	2	12	1.00	3.00	4.41	0.68	0.0500	0.4928
A(3)	GG F	0.6493	4	24	1.00	6.53	11.18	0.58	0.0500	0.3507
G(2)	GG F	0.9118	4	24	1.00	8.00	11.18	0.72	0.0500	0.0882
D(4)	GG F	0.1563	4	24	1.00	2.24	7.64	0.29	0.0500	0.8437
M(2)	GG F	0.8833	4	24	1.00	3.00	4.41	0.68	0.0500	0.1167
A(3)	GG F	0.8556	6	36	1.00	6.53	11.18	0.58	0.0500	0.1444
G(2)	GG F	0.9858	6	36	1.00	8.00	11.18	0.72	0.0500	0.0142
D(4)	GG F	0.2412	6	36	1.00	2.24	7.64	0.29	0.0500	0.7588
M(2)	GG F	0.9767	6	36	1.00	3.00	4.41	0.68	0.0500	0.0233
A(3)	GG F	0.9471	8	48	1.00	6.53	11.18	0.58	0.0500	0.0529
G(2)	GG F	0.9980	8	48	1.00	8.00	11.18	0.72	0.0500	0.0020
D(4)	GG F	0.3245	8	48	1.00	2.24	7.64	0.29	0.0500	0.6755
M(2)	GG F	0.9959	8	48	1.00	3.00	4.41	0.68	0.0500	0.0041
A(3)	GG F	0.9823	10	60	1.00	6.53	11.18	0.58	0.0500	0.0177
G(2)	GG F	0.9997	10	60	1.00	8.00	11.18	0.72	0.0500	0.0003
D(4)	GG F	0.4054	10	60	1.00	2.24	7.64	0.29	0.0500	0.5946
M(2)	GG F	0.9993	10	60	1.00	3.00	4.41	0.68	0.0500	0.0007
A(3)	GG F	1.0000	20	120	1.00	6.53	11.18	0.58	0.0500	0.0000
G(2)	GG F	1.0000	20	120	1.00	8.00	11.18	0.72	0.0500	0.0000
D(4)	GG F	0.7286	20	120	1.00	2.24	7.64	0.29	0.0500	0.2714
M(2)	GG F	1.0000	20	120	1.00	3.00	4.41	0.68	0.0500	0.0000

A(3)	GG F	1.0000	30	180	1.00	6.53	11.18	0.58	0.0500	0.0000
G(2)	GG F	1.0000	30	180	1.00	8.00	11.18	0.72	0.0500	0.0000
D(4)	GG F	0.8971	30	180	1.00	2.24	7.64	0.29	0.0500	0.1029
M(2)	GG F	1.0000	30	180	1.00	3.00	4.41	0.68	0.0500	0.0000
A(3)	GG F	1.0000	40	240	1.00	6.53	11.18	0.58	0.0500	0.0000
G(2)	GG F	1.0000	40	240	1.00	8.00	11.18	0.72	0.0500	0.0000
D(4)	GG F	0.9658	40	240	1.00	2.24	7.64	0.29	0.0500	0.0342
M(2)	GG F	1.0000	40	240	1.00	3.00	4.41	0.68	0.0500	0.0000

This report gives the power values for the various terms and sample sizes that were entered. It is much easier to consider the following plot to interpret the results.

Plots Section



From this chart, we can see that the first within-subject factor, dose level, has a power much lower than the other factors. Looking at the Sm values in the numeric table, we find that the Sm value for factor D is much less than for the other values. This explains why its power is so poor. Our options are to either increase the sample size or increase the value of Sm for factor D.

Chapter 571

Mixed Models

Introduction

The Mixed Models procedure power analyzes a variety of experimental designs in which the outcome (response) is continuous. Thus, as with the analysis of variance (ANOVA), the Mixed Models procedure is used to test hypotheses comparing various group means. Unlike ANOVA, the Mixed Models procedure relaxes the strict assumptions regarding the variances of the groups. Mixed models are very extensive. This procedure is restricted to random effects models. Random effects models are commonly used to analyze longitudinal (repeated measures) data.

The linear mixed model extends many of the classical statistical techniques, such as

- Two-sample designs (extending the t-test)
- One-way layout designs (extending one-way ANOVA)
- Factorial designs (extending factorial GLM)
- Split-plot designs (extending split-plot GLM)
- Repeated-measures designs (extending repeated-measures GLM)
- Cross-over designs (extending GLM)

Types of Linear Mixed Models

Several linear mixed model subtypes exist that are characterized by the random effects, fixed effects, and covariance structure they involve. These include fixed effects models, random effects models, and covariance pattern models.

Fixed Effects Models

A *fixed effects model* is a model where only fixed effects are included in the model. An effect (or factor) is fixed if the levels in the study represent all levels of interest of the factor, or at least all levels that are important for inference (e.g., treatment, dose, etc.). No random components are present. The general linear model is a fixed effects model. Fixed effects models can include interactions. The fixed effects can be estimated and tested using the F-test.

The fixed effects in the model include those factors for which means, standard errors, and confidence intervals will be estimated and tests of hypotheses will be performed. Other variables for which the model is to be adjusted (that are not important for estimation or hypothesis testing) may also be included in the model as fixed factors.

Random Effects Models

A *random effects model* includes both fixed and random terms in the model. An effect (or factor) is random if the levels of the factor represent a random subset of a larger group of levels (e.g., patients). The random effects are not tested, but are included to make the model more realistic.

Longitudinal Data Models

Longitudinal data arises when more than one response is measured on each subject in the study. Responses are often measured over time at fixed time points. A time point is fixed if it is pre-specified. Various variance-matrix structures can be employed to model the variance and correlation among repeated measurements.

Types of Factors

Between-Subject Factors

Between-subject factors are those that separate the experimental subjects into groups. If twelve subjects are randomly assigned to three treatment groups (four subjects per group), treatment is a between-subject factor.

Within-Subject Factors

Within-subject factors are those in which the response is measured on the same subject at several time points. Within-subject factors are those factors for which multiple levels of the factor are measured on the same subject. If each subject is measured at the low, medium, and high level of the treatment, treatment is a within-subject factor.

Technical Details

What is a Mixed Model?

In a general linear model (GLM), a random sample of the individuals in is drawn. Treatments are applied to each individual and an outcome is measured. The data so obtained are analyzed using an analysis of variance table that includes an F-test.

A mathematical model may be formulated that underlies each analysis of variance. This model expresses of the response variable as the sum of population parameters and a residual. For example, a common linear model for a two-factor experiment is

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}$$

where $i = 1, 2, \ldots, I$ (the number of levels of factor 1), $j = 1, 2, \ldots, J$ (the number of levels of factor 2), and $k = 1, 2, \ldots, K$ (the number of subjects in the study). This model expresses the value of the response variable, Y, as the sum of five components:

μ the mean.

a_i the contribution of the i^{th} level of a factor A.

b_j the contribution of the j^{th} level of a factor B.

$(ab)_{ij}$ the combined contribution (or interaction) of the i^{th} level of a factor A and the j^{th} level of a factor B.

e_{ijk} the contribution of the k^{th} individual. This is often called the *residual*.

In this example, the linear model is made up of *fixed effects* only. An effect is fixed if the levels in the study represent all levels of the factor that are of interest, or at least all levels that are important for inference (e.g., treatment, dose, etc.).

The following assumptions are made when using the F-test in a general linear model.

1. The response variable is continuous.
2. The individuals are independent.
3. The e_{ijk} follow the normal probability distribution with mean equal to zero.
4. The variances of the e_{ijk} are equal for all values of i, j, and k.

The Linear Mixed Model (LMM)

The linear mixed model (LMM) is a natural extension of the general linear model. Mixed models extend linear models by allowing for the addition of *random effects*, where the levels of the factor represent a random subset of a larger group of all possible levels (e.g., time of administration, clinic, etc.). For example, the two-factor linear model above could be augmented to include random effects such as an adjustment for each patient, since a patient may be assumed to be a random realization from a distribution of patients. The general form of the mixed model in matrix notation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where

$\mathbf{y}$ vector of responses

$\mathbf{X}$ known design matrix of the fixed effects

$\boldsymbol{\beta}$ unknown vector of fixed effects parameters to be estimated

$\mathbf{Z}$ known design matrix of the random effects

$\mathbf{u}$ unknown vector of random effects

$\boldsymbol{\varepsilon}$ unobserved vector of random errors

We assume

$\mathbf{u} \sim N(\mathbf{0},\mathbf{G})$

$\boldsymbol{\varepsilon} \sim N(\mathbf{0},\mathbf{R})$

$\mathrm{Cov}[\mathbf{u}, \boldsymbol{\varepsilon}] = \mathbf{0}$

where

$\mathbf{G}$ variance-covariance matrix of $\mathbf{u}$

$\mathbf{R}$ variance-covariance matrix of the errors $\boldsymbol{\varepsilon}$

The variance-covariance matrix of $\mathbf{y}$, denoted $\mathbf{V}$, is

$$\begin{aligned}\mathbf{V} &= \mathrm{Var}[\mathbf{y}]\\ &= \mathrm{Var}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}]\\ &= 0 + \mathrm{Var}[\mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}]\\ &= \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\end{aligned}$$

Individual Subject Formulation

Because of the size of the matrices that are involved in mixed model analysis, it is useful for computational purposes to reduce the dimensionality of the problem by analyzing the data one subject at a time. Because the data from different subjects are statistically independent, the log-likelihood of the data can be summed over the subjects, according to the formulas below. Before we look at the likelihood functions, we examine the linear mixed model for a particular subject:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, \mathrm{N}$$

where

$\mathbf{y}_i$ n_i×1 vector of responses for subject i.

$\mathbf{X}_i$ $n_i \times p$ design matrix of fixed effects for subject i (p is the number of columns in $\mathbf{X}$).

$\boldsymbol{\beta}$ p×1 vector of regression parameters.

$\mathbf{Z}_i$ $n_i \times q$ design matrix of the random effects for subject i.

$\mathbf{u}_i$ q×1 vector of random effects for subject i which has means of zero and covariance matrix $\mathbf{G}_{sub}$.

$\boldsymbol{\varepsilon}_i$ n_i×1 vector of errors for subject i with zero mean and covariance $\mathbf{R}_i$.

n_i number of repeated measurements on subject i.

N number of subjects.

The following definitions will also be useful.

$\mathbf{e}_i$ vector of residuals for subject i ($\mathbf{e}_i = \mathbf{y}_i$ - $\mathbf{X}_i\boldsymbol{\beta}$).

$\mathbf{V}_i$ $\mathrm{Var}[\mathbf{y}_i] = \mathbf{Z}_i\mathbf{G}_{sub}\mathbf{Z}_i' + \mathbf{R}_i$

To see how the individual subject mixed model formulation relates to the general form, we have

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_N \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{pmatrix}$$

In order to test the parameters in **β**, which is typically the goal in LMM analysis, the unknown parameters (**β, G**, and **R**) must be estimated. Estimates for **β** require estimates of **G** and **R**. In order to estimate **G** and **R**, the structure of **G** and **R** must be specified. Details of the specific structures for **G** and **R** are discussed later.

The following assumptions are made when using the F-test in a LMM.

1. The response variable is continuous.
2. The individuals are independent.
3. The responses follow the normal probability distribution with mean equal to zero and variance structure given by **V**.

A distinct (and arguably the most important) advantage of LMM over the GLM is flexibility in random error and random effect variance component modeling (note that the equal-variance assumption of GLM is not necessary for LMM). LMM allows you to model both heterogeneous variances and correlations among observations through the specification of the covariance matrix structures for **u** and **ε**. The variance matrix estimates are obtained using maximum likelihood (ML) or, more commonly, restricted maximum likelihood (REML). The fixed effects in the mixed model are tested using F-tests.

Structure of the Variance-Covariance Matrix

The G Matrix

The **G** matrix is the variance-covariance matrix for the random effects **u**. Typically, when the **G** matrix is used to specify the variance-covariance structure of **y**, the structure for **R** is simply $\sigma^2\mathbf{I}$. Caution should be used when both **G** and **R** are specified as complex structures, since large numbers of sometimes redundant covariance elements can result.

The **G** matrix is made up of N symmetric $\mathbf{G}_{sub}$ matrices,

$$\mathbf{G} = \begin{pmatrix} G_{sub} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & G_{sub} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & G_{sub} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & G_{sub} \end{pmatrix}$$

The dimension of $\mathbf{G}_{sub}$ is $q \times q$, where q is the number of random effects for each subject.

Structures of $\mathbf{G}_{sub}$

The structure of the $\mathbf{G}_{sub}$ matrix in this procedure is diagonal.

Diagonal $\mathbf{G}_{sub}$

$$\mathbf{G}_{sub} = \begin{pmatrix} \sigma_1{}^2 & & & \\ & \sigma_2{}^2 & & \\ & & \sigma_3{}^2 & \\ & & & \sigma_4{}^2 \end{pmatrix}$$

The R Matrix

The **R** matrix is the variance-covariance matrix for errors, **ε**. When the **R** matrix is used to specify the variance-covariance structure of **y**, the $\mathbf{G}_{sub}$ matrix is not used.

The full **R** matrix is made up of N symmetric **R** sub-matrices,

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_N \end{pmatrix}$$

where $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \cdots, \mathbf{R}_N$ are all of the same structure.

Structures of R

There are many possible structures for the sub-matrices that make up the **R** matrix. The $\mathbf{R}_{Sub}$ structures that can be specified in ***PASS*** are shown below.

Diagonal

Homogeneous

$$\begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \sigma_3^2 & \\ & & & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

Compound Symmetry

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_4\sigma_1 & \rho\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

AR(1)

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \rho^3\sigma_1\sigma_4 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho^2\sigma_2\sigma_4 \\ \rho^2\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho^3\sigma_4\sigma_1 & \rho^2\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

Toeplitz

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 & \rho_3\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 \\ \rho_3\sigma^2 & \rho_2\sigma^2 & \rho_1\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_2\sigma_1\sigma_3 & \rho_3\sigma_1\sigma_4 \\ \rho_1\sigma_2\sigma_1 & \sigma_2^2 & \rho_1\sigma_2\sigma_3 & \rho_2\sigma_2\sigma_4 \\ \rho_2\sigma_3\sigma_1 & \rho_1\sigma_3\sigma_2 & \sigma_3^2 & \rho_1\sigma_3\sigma_4 \\ \rho_3\sigma_4\sigma_1 & \rho_2\sigma_4\sigma_2 & \rho_1\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

Toeplitz(2)

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho_1\sigma^2 & & \\ \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \\ & \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 \\ & & \rho_1\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & & \\ \rho_1\sigma_2\sigma_1 & \sigma_2^2 & \rho_1\sigma_2\sigma_3 & \\ & \rho_1\sigma_3\sigma_2 & \sigma_3^2 & \rho_1\sigma_3\sigma_4 \\ & & \rho_1\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho_1 & & \\ \rho_1 & 1 & \rho_1 & \\ & \rho_1 & 1 & \rho_1 \\ & & \rho_1 & 1 \end{pmatrix}$$

Banded(2)

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & & \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \\ & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ & & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & & \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \\ & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ & & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho & & \\ \rho & 1 & \rho & \\ & \rho & 1 & \rho \\ & & \rho & 1 \end{pmatrix}$$

Note: This is the same as Toeplitz(1).

Banded(3)

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ & \rho\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho & \rho & \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ & \rho & \rho & 1 \end{pmatrix}$$

Unstructured

Homogeneous

$$\begin{pmatrix} \sigma^2 & \rho_{12}\sigma^2 & \rho_{13}\sigma^2 & \rho_{14}\sigma^2 \\ \rho_{21}\sigma^2 & \sigma^2 & \rho_{23}\sigma^2 & \rho_{24}\sigma^2 \\ \rho_{31}\sigma^2 & \rho_{32}\sigma^2 & \sigma^2 & \rho_{34}\sigma^2 \\ \rho_{41}\sigma^2 & \rho_{42}\sigma^2 & \rho_{43}\sigma^2 & \sigma^2 \end{pmatrix}$$

Heterogeneous

$$\begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{31}\sigma_3\sigma_1 & \rho_{32}\sigma_3\sigma_2 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{41}\sigma_4\sigma_1 & \rho_{42}\sigma_4\sigma_2 & \rho_{43}\sigma_4\sigma_3 & \sigma_4^2 \end{pmatrix}$$

Correlation

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$$

Partitioning the Variance-Covariance Structure with Groups

In the case where it is expected that the variance-covariance parameters are different across groups of a between-subjects factor, a different set of **R** or **G** parameters can be specified for each group. This produces a set of variance-covariance parameters that is different for each level of the chosen group variable, but each set has the same structure.

Likelihood Formulas

There are two types of likelihood estimation methods that are generally considered in mixed model estimation: maximum likelihood (ML) and restricted maximum likelihood (REML). REML is generally favored over ML because the variance estimates using REML are unbiased for small sample sizes, whereas ML estimates are unbiased only asymptotically (see Littell et al., 2006 or Demidenko, 2004). Both estimation methods are available in ***PASS***.

Maximum Likelihood

The general form -2 log-likelihood ML function is

$$-2L_{ML}(\boldsymbol{\beta},\mathbf{G},\mathbf{R}) = \ln|\mathbf{V}| + \mathbf{e}'\mathbf{V}^{-1}\mathbf{e} + N_T\ln(2\pi)$$

The equivalent individual subject form is

$$-2L_{\mathbf{ML}}(\boldsymbol{\beta},\mathbf{G},\mathbf{R})=\sum_{\mathbf{i}=1}^{N}\left(\ln|\mathbf{V_i}|+\mathbf{e_i'V_i^{-1}e_i}\right)+\mathbf{N_T}\ln(2\pi)$$

where N_T is the total number of observations, or

$$N_T=\sum_{i=1}^{N}n_i$$

Restricted Maximum Likelihood

The general form -2 log-likelihood REML function is

$$-2L_{REML}(\boldsymbol{\beta},\mathbf{G},\mathbf{R})=\ln|\mathbf{V}|+\mathbf{e'V^{-1}e}+\ln|\mathbf{X'V^{-1}X}|+(N_T-p)\ln(2\pi)$$

The equivalent individual subject form is

$$-2L_{REML}(\boldsymbol{\beta},\mathbf{G},\mathbf{R})=\sum_{i=1}^{N}\left[\ln|\mathbf{V}_i|+\mathbf{e}_i'\mathbf{V}_i^{-1}\mathbf{e}_i\right]+\ln\left|\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right|+(N_T-p)\ln(2\pi)$$

where, again, N_T is the total number of observations, or

$$N_T=\sum_{i=1}^{N}n_i$$

and p is the number of columns in **X** or $\mathbf{X}_i$.

Estimating and Testing Fixed Effects Parameters

The estimation phase in the analysis of a mixed model produces variance and covariance parameter estimates of the elements of **G** and **R**, giving $\hat{\mathbf{R}}$ and $\hat{\mathbf{G}}$, and hence, $\hat{\mathbf{V}}$. The REML and ML solutions for $\hat{\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X'\hat{V}^{-1}X}\right)^{-1}\mathbf{X'\hat{V}^{-1}y}$$

with estimated variance-covariance

$$\hat{\Sigma}=\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right)=\left(\mathbf{X'\hat{V}^{-1}X}\right)^{-1}$$

See, for example, Brown and Prescott (2006), Muller and Stewart (2006), or Demidenko (2004) for more details of the estimating equations.

Hypothesis tests and confidence intervals for **β** are formed using a linear combination matrix (or vector) **L**. Although you don’t have to specify **L**, it is important that you understand how its function.

L Matrix Details

L matrices specify linear combinations of **β** corresponding to means or hypothesis tests of interest. Essentially, the **L** matrix defines the mean or test. The number of columns in each **L** matrix is the same as the number of elements of **β**. For estimating a particular mean, the **L** matrix consists of a single row. For hypothesis tests, the number of rows of **L** varies according to the test. Below are some examples of **L** matrices that arise in common analyses:

L Matrix for Testing a Single Factor (Food with 4 levels) in a Single-Factor Model

No.	Effect	Food	L1	L2	L3
1	Intercept				
2	Food	HighIron	1.0000	1.0000	1.0000
3	Food	LowIron	-1.0000		
4	Food	None		-1.0000	
5	Food	Salicyl			-1.0000

L Matrix for a Single Mean (LowIron) of a Single Factor (4 levels) in a Single-Factor Model

No.	Effect	Food	L1
1	Intercept		1.0000
2	Food	HighIron	
3	Food	LowIron	1.0000
4	Food	None	
5	Food	Salicyl	

L Matrix for Testing a Single Factor (Drug – 3 levels) in a Two-Factor Model with Interaction

No.	Effect	Drug	Time	L1	L2
1	Intercept				
2	Drug	Kerlosin		1.0000	1.0000
3	Drug	Laposec		-1.0000	
4	Drug	Placebo			-1.0000
5	Time		0.5		
6	Time		1		
7	Time		1.5		
8	Time		2		
9	Time		2.5		
10	Time		3		
11	Drug*Time	Kerlosin	0.5	0.1667	0.1667
12	Drug*Time	Kerlosin	1	0.1667	0.1667
13	Drug*Time	Kerlosin	1.5	0.1667	0.1667
14	Drug*Time	Kerlosin	2	0.1667	0.1667
15	Drug*Time	Kerlosin	2.5	0.1667	0.1667
16	Drug*Time	Kerlosin	3	0.1667	0.1667
17	Drug*Time	Laposec	0.5	-0.1667	
18	Drug*Time	Laposec	1	-0.1667	
19	Drug*Time	Laposec	1.5	-0.1667	
20	Drug*Time	Laposec	2	-0.1667	
21	Drug*Time	Laposec	2.5	-0.1667	
22	Drug*Time	Laposec	3	-0.1667	
23	Drug*Time	Placebo	0.5		-0.1667
24	Drug*Time	Placebo	1		-0.1667
25	Drug*Time	Placebo	1.5		-0.1667
26	Drug*Time	Placebo	2		-0.1667
27	Drug*Time	Placebo	2.5		-0.1667
28	Drug*Time	Placebo	3		-0.1667

Kenward and Roger Fixed Effects Hypothesis Tests

Hypothesis tests have the general form

$$H_0\text{: } \mathbf{L\beta} = \mathbf{0}$$

where **L** is a linear contrast matrix of rank *h* corresponding to the desired comparisons to be made in the hypothesis test. Let *d* be the denominator degrees of freedom and *q* be the number of variance-covariance parameters, which is the dimension of **W** (defined below).

The Kenward and Roger (1997) test statistic for testing H_0 is currently the most recommended method of specifying the F-ratio and its degrees of freedom. The F-ratio is

$$\mathbf{F}_{\mathbf{h,d}} = \frac{\lambda}{h}\hat{\boldsymbol{\beta}}'\mathbf{L}'\left(\mathbf{LC}*\mathbf{L}'\right)^{-1}\mathbf{L}\hat{\boldsymbol{\beta}}$$

where

$$\mathbf{C}* = \mathbf{C} + 2\mathbf{C}\left\{\sum_{\mathbf{r}=1}^{\mathbf{q}}\sum_{\mathbf{s}=1}^{\mathbf{q}}\mathbf{W}_{\mathbf{rs}}\left(\mathbf{Q}_{\mathbf{rs}} - \mathbf{P}_{\mathbf{r}}\mathbf{CP}_{\mathbf{s}} - \frac{1}{4}\mathbf{S}_{\mathbf{rs}}\right)\right\}\mathbf{C}$$

$$\mathbf{C} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

$$\mathbf{Q}_{\mathbf{rs}} = \mathbf{X}'\mathbf{V}^{-1}\dot{\mathbf{V}}_{\mathbf{r}}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\mathbf{s}}\mathbf{V}^{-1}\mathbf{X} = \sum_{i=1}^{N}\mathbf{X}_{\mathbf{i}}'\mathbf{V}_{\mathbf{i}}^{-1}\dot{\mathbf{V}}_{\mathbf{ri}}\mathbf{V}_{\mathbf{i}}^{-1}\dot{\mathbf{V}}_{\mathbf{si}}\mathbf{V}_{\mathbf{i}}^{-1}\mathbf{X}_{\mathbf{i}}$$

$$\mathbf{P}_{\mathbf{r}} = -\mathbf{X}'\mathbf{V}^{-1}\dot{\mathbf{V}}_{\mathbf{r}}\mathbf{V}^{-1}\mathbf{X} = -\sum_{i=1}^{N}\mathbf{X}_{\mathbf{i}}'\mathbf{V}_{\mathbf{i}}^{-1}\dot{\mathbf{V}}_{\mathbf{ri}}\mathbf{V}_{\mathbf{i}}^{-1}\mathbf{X}_{\mathbf{i}}$$

$$\mathbf{S}_{\mathbf{rs}} = \mathbf{X}'\mathbf{V}^{-1}\ddot{\mathbf{V}}_{\mathbf{rs}}\mathbf{V}^{-1}\mathbf{X} = \sum_{i=1}^{N}\mathbf{X}_{\mathbf{i}}'\mathbf{V}_{\mathbf{i}}^{-1}\ddot{\mathbf{V}}_{\mathbf{rsi}}\mathbf{V}_{\mathbf{i}}^{-1}\mathbf{X}_{\mathbf{i}}$$

$$\mathbf{W} = \mathbf{H}^{-1}$$

$$\{\mathbf{H}\}_{\mathbf{rs}} = \{\text{Hessian}\}_{\mathbf{rs}}$$

$$\dot{\mathbf{V}}_{\mathbf{r}} = \frac{\partial\mathbf{V}}{\partial\sigma_{\mathbf{r}}}$$

$$\ddot{\mathbf{V}}_{\mathbf{rs}} = \frac{\partial^2\mathbf{V}}{\partial\sigma_{\mathbf{r}}\partial\sigma_{\mathbf{s}}}$$

$$\mathbf{T} = \mathbf{L}'\left(\mathbf{LCL}'\right)^{-1}\mathbf{L}$$

$$a_1 = \sum_{\mathbf{r}=1}^{\mathbf{q}}\sum_{\mathbf{s}=1}^{\mathbf{q}}\mathbf{W}_{\mathbf{rs}}\,\mathrm{tr}(\mathbf{TCP}_{\mathbf{r}}\mathbf{C})\,\mathrm{tr}(\mathbf{TCP}_{\mathbf{s}}\mathbf{C}),\quad a_2 = \sum_{\mathbf{r}=1}^{\mathbf{q}}\sum_{\mathbf{s}=1}^{\mathbf{q}}\mathbf{W}_{\mathbf{rs}}\,\mathrm{tr}(\mathbf{TCP}_{\mathbf{r}}\mathbf{CTCP}_{\mathbf{s}}\mathbf{C})$$

$$a_3 = \frac{a_1 + 6a_2}{2h},\quad e = \left(1 - \frac{a_2}{h}\right)^{-1},\quad v = \frac{2}{h}\left\{\frac{1 + c_1a_3}{\left(1 - c_2a_3\right)^2\left(1 - c_3a_3\right)}\right\}$$

$$c_1 = \frac{g}{3h + 2(1-g)},\ c_2 = \frac{h-g}{3h + 2(1-g)},\ c_3 = \frac{h+2-g}{3h + 2(1-g)},\ c_4 = \frac{v}{2e^2}$$

$$g = \frac{(h+1)a_1 - (h+4)a_2}{(h+2)a_2}$$

$$d = 4 + \frac{h+2}{c_4 h - 1},\ \lambda = \frac{d}{e(d\text{-}2)}$$

Solution Algorithms

Methods for Finding Likelihood Solutions (Newton-Raphson, Fisher Scoring, MIVQUE, and Differential Evolution)

There are four techniques in the Mixed Models procedure for determining the maximum likelihood or restricted maximum likelihood solution (optimum): Newton-Raphson, Fisher Scoring, MIVQUE, and Differential Evolution.

The general steps for the Newton-Raphson, Fisher Scoring, and Differential Evolution techniques are (let θ be the overall covariance parameter vector):

1. Roughly estimate θ according to the specified structure for each.
2. Evaluate the likelihood of the model given the data and the estimates of θ.
3. Improve upon the estimates of θ using a search algorithm. (Improvement is defined as an increase in likelihood.)
4. Iterate until maximum likelihood is reached, according to some convergence criterion.
5. Use the final θ estimates to estimate $\boldsymbol{\beta}$.

Newton-Raphson and Fisher Scoring

The differences in the techniques revolve around the initial estimates in Step 1, and the improvements in estimates made in Step 3. For the Newton-Raphson and Fisher Scoring techniques, Step 3 occurs as follows:

3a. With the estimated θ, compute the gradient vector **g**, and the Hessian matrix **H**.
3b. Compute $\mathbf{d} = -\mathbf{H}^{-1}\mathbf{g}$.
3c. Let $\lambda = 1$.
3d. Compute new estimates for θ, iteratively, using $\theta_i = \theta_{i\text{-}1} + \lambda\mathbf{d}$.
3e. If θ_i is a valid set of covariance parameters and improves the likelihood, continue to 3f. Otherwise, reduce λ by half and return to Step 3d.
3f. Check for convergence. If the convergence criteria (small change in -2log-likelihood) are met, stop. If the convergence criteria are not met, go back to Step 3a.

The gradient vector **g**, and the Hessian matrix **H**, used for the Newton-Raphson and Fisher Scoring techniques for solving the REML equations are shown in the following table:

REML Gradient (g) and Hessian (H)

Technique	Gradient (g)	Hessian (H)
Newton-Raphson	$\mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3$	$\mathbf{H}_1 + \mathbf{H}_2 + \mathbf{H}_3$
Fisher Scoring	$\mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3$	$-\mathbf{H}_1 + \mathbf{H}_3$

The gradient vector **g**, and the Hessian matrix **H**, used for the Newton-Raphson and Fisher Scoring techniques for solving the ML equations are shown in the following table:

ML Gradient (g) and Hessian (H)

Technique	Gradient (g)	Hessian (H)
Newton-Raphson	$\mathbf{g}_1 + \mathbf{g}_2$	$\mathbf{H}_1 + \mathbf{H}_2$
Fisher Scoring	$\mathbf{g}_1 + \mathbf{g}_2$	$-\mathbf{H}_1$

where $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ are defined as in Wolfinger, Tobias, and Sall (1994).

Definitons

$$\dot{\mathbf{V}}_{\mathbf{ri}} = \frac{\partial \mathbf{V}_\mathbf{i}}{\partial \boldsymbol{\sigma}_\mathbf{r}}, \quad \ddot{\mathbf{V}}_{\mathbf{rsi}} = \frac{\partial^2 \mathbf{V}_\mathbf{i}}{\partial \boldsymbol{\sigma}_\mathbf{r} \partial \boldsymbol{\sigma}_\mathbf{s}}, \mathbf{e}_\mathbf{i} = \mathbf{y}_\mathbf{i} - \mathbf{X}_\mathbf{i}\boldsymbol{\beta}, \quad \mathbf{A}_\mathbf{i} = \mathbf{X}_\mathbf{i}'\mathbf{V}_\mathbf{i}^{-1}\mathbf{X}_\mathbf{i}, \ \mathbf{A} = \sum_{i=1}^{N} \mathbf{X}_\mathbf{i}'\mathbf{V}_\mathbf{i}^{-1}\mathbf{X}_\mathbf{i} = \sum_{i=1}^{N} \mathbf{A}_\mathbf{i},$$

$$\mathbf{C} = \mathbf{A}^{-1}, \quad \dot{\mathbf{A}}_\mathbf{r} = \sum \mathbf{X}_\mathbf{i}'\left(\frac{\partial \mathbf{V}_\mathbf{i}^{-1}}{\partial \boldsymbol{\sigma}_\mathbf{r}}\right)\mathbf{X}_\mathbf{i} = -\sum \mathbf{X}_\mathbf{i}'\left(\mathbf{V}_\mathbf{i}^{-1}\dot{\mathbf{V}}_{\mathbf{ri}}\mathbf{V}_\mathbf{i}^{-1}\right)\mathbf{X}_\mathbf{i} = -\mathbf{P}_\mathbf{r}$$

$$\mathbf{X}^* = \mathbf{X}\mathbf{K}, \quad \mathbf{K}\mathbf{K}' = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

Likelihoods

$$\mathbf{l}_1 = \frac{1}{2}\sum_{i=1}^{N} \ln\left|\mathbf{V}_\mathbf{i}\right|, \quad \mathbf{l}_2 = \frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_\mathbf{i}'\mathbf{V}_\mathbf{i}^{-1}\mathbf{e}_\mathbf{i}, \ \mathbf{l}_3 = \frac{1}{2}\ln\left|\sum_{i=1}^{N} \mathbf{X}_\mathbf{i}'\mathbf{V}_\mathbf{i}^{-1}\mathbf{X}_\mathbf{i}\right| = \frac{1}{2}\ln\left|\sum_{i=1}^{N} \mathbf{A}_\mathbf{i}\right| = \frac{1}{2}\ln\left|\mathbf{A}\right|$$

First Derivatives

$$\mathbf{g}_{1\mathbf{r}} = \frac{\partial \mathbf{l}_1}{\partial \boldsymbol{\sigma}_\mathbf{r}} = \frac{1}{2}\sum_{i=1}^{N} \operatorname{tr}\left(\mathbf{V}_\mathbf{i}^{-1}\dot{\mathbf{V}}_{\mathbf{ri}}\right)$$

$$\mathbf{g}_{2\mathbf{r}} = \frac{\partial \mathbf{l}_2}{\partial \boldsymbol{\sigma}_\mathbf{r}} = -\frac{1}{2}\sum_{i=1}^{N} \mathbf{e}_\mathbf{i}'\mathbf{V}_\mathbf{i}^{-1}\dot{\mathbf{V}}_{\mathbf{ri}}\mathbf{V}_\mathbf{i}^{-1}\mathbf{e}_\mathbf{i}$$

$$\mathbf{g}_{3\mathbf{r}} = \frac{\partial \mathbf{l}_3}{\partial \boldsymbol{\sigma}_\mathbf{r}} = -\frac{1}{2}\operatorname{tr}\left[\mathbf{H}_3^\mathbf{r}\right]$$

Second Derivatives

$$\mathbf{H}_{\mathbf{1rs}} = \frac{\partial^2 \mathbf{l}_1}{\partial \boldsymbol{\sigma}_{\mathbf{r}} \partial \boldsymbol{\sigma}_{\mathbf{s}}} = -\frac{1}{2}\sum_{i=1}^{N}\left\{\operatorname{tr}\left(\mathbf{V}_{\mathbf{i}}^{-1}\ddot{\mathbf{V}}_{\mathbf{rsi}}\right) - \operatorname{tr}\left(\mathbf{V}_{\mathbf{i}}^{-1}\dot{\mathbf{V}}_{\mathbf{ri}}\mathbf{V}_{\mathbf{i}}^{-1}\dot{\mathbf{V}}_{\mathbf{si}}\right)\right\}$$

$$\mathbf{H}_{\mathbf{2rs}} = \frac{\partial^2 \mathbf{l}_2}{\partial \boldsymbol{\sigma}_{\mathbf{r}} \partial \boldsymbol{\sigma}_{\mathbf{s}}} = \frac{1}{2}\left(\mathbf{H}_{\mathbf{2}}^{\mathbf{rs}} - 2\mathbf{H}_{\mathbf{2}}^{\mathbf{r'}}\mathbf{H}_{\mathbf{2}}^{\mathbf{s}}\right)$$

$$\mathbf{H}_{\mathbf{3rs}} = \frac{\partial^2 \mathbf{l}_3}{\partial \boldsymbol{\sigma}_{\mathbf{r}} \partial \boldsymbol{\sigma}_{\mathbf{s}}} = \frac{1}{2}\operatorname{tr}\left(\mathbf{H}_{\mathbf{3}}^{\mathbf{rs}} - \mathbf{H}_{\mathbf{3}}^{\mathbf{r}}\mathbf{H}_{\mathbf{3}}^{\mathbf{s}}\right)$$

See Wolfinger, Tobias, and Sall (1994), page 1299, for details.

MIVQUE

The MIVQUE estimates of $\boldsymbol{\theta}$ in REML estimation are found by solving

$$-(\mathbf{H}_1 + \mathbf{H}_3)\boldsymbol{\theta} = -\mathbf{g}_2 .$$

The MIVQUE estimates of $\boldsymbol{\theta}$ in ML estimation are found by solving

$$-\mathbf{H}_1\boldsymbol{\theta} = -\mathbf{g}_2 .$$

See Wolfinger, Tobias, and Sall (1994), page 1306, for details.

Differential Evolution

The differential evolution techniques used in this procedure for the ML and REML optimization are described in Price, Storn, and Lampinen (2005). This algorithm is very slow, but it is also very robust. As it stands now, it is too slow to be used. However, as computers become faster, this algorithm will become more viable.

Specifying the Minimum Detectable Difference

The four main parameters of a power analysis are the sample size, the effect size, the significance level, and the power level. Other extraneous parameters, such as the variance, must also be specified. This section describes the specification of the effect size, or minimum detectable difference (MDD) as we choose to call it in this chapter.

Power is defined as the probability of rejecting the null hypothesis of zero difference when the actual difference is a given amount. As the size of the actual difference increases, so does the power. The MDD (or minimum effect size) is the smallest difference among the population means that will be detected by an experiment at the specified settings of the other parameters.

Typically, a longitudinal design includes a between-subjects factor, a within-subject factor, and their interaction. A MDD must be specified for each. As the number of factors grows, the number of interactions grows, and the number of MDD’s that must be specified also grows. It becomes crucial that you specify these values in a meaningful and accurate way.

In the Repeated Measures module, PASS only requires the standard deviation of the group means. Unfortunately, this is a quantity that researchers have very little experience with. It seldom appears on any of the standard reports that are produced by commercial software. It is seldom present in the written reports of analyses. A different method is used to specify the MDD in *PASS*.

In this routine, the MDD is specified as the difference between the smallest and largest effects. For example, suppose that a factor has three levels with means 10, 15, 18. The detectable difference is 18 – 10 = 8. This is a simply quantity that is easy to interpret.

The Effect Pattern for Factors

When there are more than two means, the minimum detectable difference does not uniquely define a set of means that can be simulated. For example the following sets of means all have identical MDD's, but the means themselves are quite different: (10, 12, 18), (10, 14, 18), and (10, 18, 18). The following method for defining the pattern is quite informative:

1. Set the first (low) value to -0.5.
2. Set the last (high) value to 0.5.
3. Set each value in between to (Mean – Min) / MDD – 0.5.

Using these steps, the three sets of means may be reduced to MDD and a pattern as follows:

Original	MDD	Pattern	MDD x Pattern
10, 12, 18	8	-0.5, -0.25, 0.5	-4, -2, 4
10, 14, 18	8	-0.5, 0.0, 0.5	-4, 0, 4
10, 18, 18	8	-0.5, 0.5, 0.5	-4, 4, 4

Thus, each set of means can be easily reduced to two components: the MDD and a pattern. This is the method that PASS uses to supply the sets of means. Using this method, it is easy to compare various sizes of means. You simply enter different values for the MDD, keeping the pattern the same.

The Effect Pattern for Interactions

Specifying the structure of the interactions is a little more problematic. Often, you are not interested in the interaction. When the interaction is of interest, you may have only a vague idea of its structure. Part of your analysis will be to investigate the effect of the interaction, with little or no knowledge of its pattern beforehand.

Specifying the MDD for the interaction is somewhat intuitive. The MDD defines the largest difference among the interaction effects. A difficulty still arises in that there is a very large number of possible patterns, many of which are useful. For planning purposes, we have decided to use a standard pattern in *PASS*. The interaction pattern used in *PASS* is defined as the Kronecker product of the factor patterns that make up the interaction, scaled so that the largest value is 0.5 and the smallest value is -0.5.

Example

Suppose that a two-factor interaction is made up of a three-level factor A with a pattern of -0.5, 0, 0.5 and a two-level factor B with a pattern of -0.5, 0.5. The interaction pattern would be found as follows. The Kronecker product of these two patterns is 0.25, -0.25, 0.0, 0.0, -0.25, 0.25. Rescaling so that the minimum is -0.5 and the maximum is 0.5 is achieved by doubling the values. The final interaction pattern is 0.5, -0.5, 0.0, 0.0, -0.5, 0.5. This pattern compares the difference due to factor B across the levels of factor A.

Suppose the MDD for A is set to 8, the MDD for B is set to 15, and the MDD for AB is set to 10. Suppose the overall mean is 12. The six cell means would be found by adding the individual effects as follows

Term	A1B1	A1B2	A2B1	A2B2	A3B1	A3B2
A	-4.0	-4.0	0.0	0.0	4.0	4.0
B	-7.5	7.5	-7.5	7.5	-7.5	7.5
AB	5.0	-5.0	0.0	0.0	-5.0	5.0
Total	-6.5	-1.5	-7.5	7.5	-8.5	16.5
Overall	12.0	12.0	12.0	12.0	12.0	12.0
Cell Mean	5.5	10.5	4.5	19.5	3.5	28.5

The cell means are then used to simulate the data for power and sample size calculation. This set of cell means has the MDD's and patterns specified.

Specifying the Simulated Variance-Covariance Matrix

As stated above, the variance of **y** is **V** = **ZGZ'** + **R**. In this *PASS* Mixed Models module, **ZGZ'** is called the *random component* and **R** is called the *residual component*. Since **V** is block-diagonal (with one block for each subject), it is specified by specifying the random and residual components for one subject and repeating those components for each subject.

ZGZ'

When the random component is included, the model is called a *random effects model.* For a design with four time points, the structure of **ZGZ'** is

$$\begin{pmatrix} g & g & g & g \\ g & g & g & g \\ g & g & g & g \\ g & g & g & g \end{pmatrix} = g[\mathbf{J}_4]$$

This structure only requires a single value: *g*.

R

The structure of **R** is quite flexible. Since it is a variance-covariance matrix, the only stipulation is that it must be non-negative definite. Possible choices for **R** when the variance is constant are

Diagonal

$$\begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \sigma^2 \end{pmatrix}$$

Constant (Compound Symetric)

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

AR(1)

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

List

$$\begin{pmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 & \rho_3\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 \\ \rho_3\sigma^2 & \rho_2\sigma^2 & \rho_1\sigma^2 & \sigma^2 \end{pmatrix}$$

Possible choices for R when the variance is allowed to vary are

Diagonal

$$\begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \sigma_3^2 & \\ & & & \sigma_4^2 \end{pmatrix}$$

Constant

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma^2 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix}$$

AR(1)

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \rho^3\sigma_1\sigma_4 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho^2\sigma_2\sigma_4 \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho^3\sigma_1\sigma_4 & \rho^2\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix}$$

List

$$\begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_2\sigma_1\sigma_3 & \rho_3\sigma_1\sigma_4 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 & \rho_1\sigma_2\sigma_3 & \rho_2\sigma_2\sigma_4 \\ \rho_2\sigma_1\sigma_3 & \rho_1\sigma_2\sigma_3 & \sigma_3^2 & \rho_1\sigma_3\sigma_4 \\ \rho_3\sigma_1\sigma_4 & \rho_2\sigma_2\sigma_4 & \rho_1\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix}$$

ZGZ' + R

When **ZGZ'** is included and **R** is set to diagonal as recommended, the value of **V** becomes

$$\begin{pmatrix} \sigma^2 + g & g & g & g \\ g & \sigma^2 + g & g & g \\ g & g & \sigma^2 + g & g \\ g & g & g & \sigma^2 + g \end{pmatrix}$$

This is the compound symmetric pattern that is assumed in the repeated measures analysis of variance (RMANOVA). This model is often used to compare LMM with RMANOVA.

Power Calculations using Computer Simulation

Computer simulation allows us to estimate the power that is actually achieved by a test procedure in situations such as LMM that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time. The simulations can still be time consuming, so we have proposed some steps below that will significantly shortened then time needed to obtain answers.

The Simulation

It is important that you understand how the simulation is setup. There are three main tabs (panels or windows) that contain parameters that you will need to set. These are the Data tab, the Covariance tab, and the Fitted Model tab.

Data Tab

The Data tab contains all the parameters associated with the sample size, the effect size, and the significance level (alpha). The effect size parameters define the experimental design.

Covariance Tab

The Covariance tab specifies the parameters used to define the covariance of the data that is generated. Note that the covariance of the data you generate does not have to match the covariance model that you use to fit the data. In fact, since you seldom know even the structure of the true covariance matrix, it is more realistic to generate data using one type of covariance matrix and then fit the generated data with a different covariance structure.

Fitted Model Tab

The Fitted Model tab specifies the covariance matrix that is actually fit to your data. As stated above, the model does not need to coincide with the model used to generate the data. It may be more realistic if it does not.

Steps in Conducting a Simulation Analysis

Simulation Steps

The steps to a simulation study are

1. Specify the design that will be studied. Enter the sample size, MDD's, and variance covariance matrix. Specify the covariance matrix of the model that is fitted to the data.

2. Generate random samples from the design specified. Calculate the F-tests from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's power.

3. Repeat step 2 several hundred or more times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection.

4. Additionally, you can run a separate simulation to determine if the significance level (alpha) of the F-test matches the significance level you have selected. This is done by setting the MDD's to zero.

Saving Simulation Time

Simulations for large models or large sample sizes take several hours to run. The simulation time can be reduced by running the simulation in two steps.

1. Specify a reasonable range for the group sample sizes. For example, you may want to try group sample sizes of 5, 10, 30, 80, and 120. Set the number of simulations to 300 or 500. Although these is not a large enough simulation size to give you definitive results, you can study the confidence intervals for power provided in the reports and plots to determine a reduced range of sample sizes.
2. Reduce the range of the sample size values, increase the number of simulations to 1000 or 2000, and rerun the simulations. These simulations may run for a while, so be prepared for running times of several minutes or hours. The power values that come from these simulations should be very precise.

Generating the Random Numbers

The simulation proceeds by generating the normal random deviates in groups that are the size of the number of time points. That is, a set of normals are generated for a single subject. This set of normals is transformed into a set of normals having the desired covariance structure using the commonly know technique of multiplying the generated unit normals by the square root of the variance-covariance matrix. The square root is taken using the Choleski decompositions. The resulting response vector matrix has the desired covariance matrix.

Symbolically, suppose there are t time points and further suppose that the desired variance-covariance matrix of the data to be simulated is given by $\mathbf{V}$. Find a matrix $\mathbf{W}$ such that $\mathbf{V} = \mathbf{WW'}$. Note that, by construction, W is lower triangular. If we generate t unit random normal deviates and place them in a vector $\mathbf{z}$, the vector $\mathbf{y} = \mathbf{Wz}$ has variance-covariance matrix $\mathbf{WW'} = \mathbf{V}$. Finally, the appropriate cell mean (based on the minimum detectable differences) is added to the $\mathbf{y}$ to obtain the simulation data with the desired properties. Once all of the data required for a complete experiment is generated, the mixed model is solved using the mixed model algorithm coded for ***NCSS's*** mixed model procedure.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options necessary to define the sample size, significance level, effect size, and simulation size.

Solve For

Find (Solve For)

This procedure always solves for the power, so there is no specific Solve For option.

Sample Size

n (Subjects Per Group)

Specify one or more values for the number of subjects per group. The total sample size is the sum of the individual group sizes across all groups.

You can specify a list such as *2 4 6*. The items in the list may be separated with commas or blanks. The interpretation of the list depends on the =n's check box. When the =n's box is checked, a separate analysis is calculated for each value of *n*. When the =n's box is not checked, ***PASS*** uses the *n's* as the actual group sizes. In this case, the number of items entered must match the number of groups in the design, which is equal to the product of the number of levels of A and B.

You can also enter the sample sizes in columns of the spreadsheet. The column contains the group sample sizes, one per row. Columns are indicated by adding an equals sign to the left of the first entry. For example, if you have entered a set of unequal group sample sizes in column 2, you would enter '=C2' here. Multiple columns, such as columns 2 through 4, are specified as '=C2 C3 C4'.

= n's

This option controls whether or not the number of subjects per group is to be equal for all groups. When checked, the number of subjects per group is equal for all groups. A list of values such as '5 10 15' represents three designs: one with five per group, one with ten per group, and one with fifteen per group. A simulation is conducted for each value.

When this option is not checked, the n's are assumed to be unequal. A list of values represents the size of the individual groups. For example, '5 10 15' represents a single, three-group design with five in the first group, ten in the second group, and fifteen in the third group. If four values are needed, but only two are entered, the last value is carried forward.

Error Rates

Alpha

This option specifies the probability of a type-I error (alpha) for each factor and interaction. A type-I error occurs when you reject the null hypothesis of zero effects when in fact they are zero. Since they are probabilities, alpha values must be between zero and one. Routinely, the value of 0.05 is used for alpha. This value may be interpreted as meaning that about one F-test in twenty will falsely reject the null hypothesis.

Simulations

Simulations

This option specifies the number of iterations, M, used in each simulation. Larger numbers of iterations result in longer run times but more accurate results.

The precision of the simulated power estimates can be determined by recognizing that they follow the binomial distribution. Thus, confidence intervals may be constructed for power estimates. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the 'Precision' amount given in the table.

Simulation Size M	Precision when Power = 0.50	Precision when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.014 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional precision achieved.

Because of the long run time needed to obtain an estimate of power, it is crucial that you set this parameter carefully. We suggest the following two-step procedure.

Step 1

Set the number of simulations to 200, 300, or 500 and run examples for a realistic range of sample sizes. Study the results (especially the confidence intervals) to determine a range of sample sizes you want to investigate more carefully.

Step 2

Reduce the range of N to just one or two values and set the number of simulations to a large amount.

Step 3 (Optional)

Take the rest of the day off—the simulation may take awhile!

Effect Size

Specify Effects Using

Indicate which options are used to specify the Effect Size. The possible choices are Means in Spreadsheet Columns or the reset of the options on this panel.

Means in Spreadsheet Columns

Specify spreadsheet columns containing a hypothesized means matrix that represents the minimum detectable differences among the means. Under the null hypothesis, this matrix is all zeros.

The between-subject factors (A & B) are represented across the columns of the spreadsheet and the within-subject factors (C & D) are represented down the rows. The number of columns specified must equal the number of groups, which is equal to the number of levels in A times the number of levels in B. The number of rows must equal the number of time points, which is equal to the number of levels in C multiplied by the number of levels in D.

For example, suppose you are designing an experiment that is to have two between factors (A & B) and two within factors (C & D). Suppose each of the four factors has two levels. The columns of the spreadsheet would represent

A1B1 A1B2 A2B1 A2B2.

The rows of the spreadsheet would represent

C1D1

C1D2

C2D1

C2D2

Example

To see how this option works, consider the following table of hypothesized means for an experiment with one between factor (A) having two groups and one within factor (C) having three time periods. The values in columns C1 and C2 of the spreadsheet are

C1	C2
2.0	4.0
4.0	6.0
6.0	11.0

By subtracting the appropriate means, the following table of effects results

	C1	C2	Means	Effects
Row1	0.5	-0.5	3.0	-2.5
Row2	0.5	-0.5	5.0	-0.5
Row3	-1.0	1.0	8.5	3.0
	---	---	---	
Means	4.0	7.0	5.5	
Effects	-1.5	1.5		

Effect Size – Factors Separating Subjects into Groups (Between)

These options specify the effect sizes of the between-subjects factors (A and B).

Levels

Specify the number of levels (categories) in this factor. Typical values are from 2 to 8. Set this to a blank (or 0) to ignore the factor in the design.

Effects Pattern

This option specifies the pattern of the means for this factor. This pattern is multiplied by the Detectable Difference to form the factor means used in the simulation. For example, suppose that the pattern of a five-category factor is -.50 -.25 0.0 .25 0.5 and the detectable difference is 10. The resulting effects are -5.0 -2.5 0.0 2.5 5.0. The power reported by the simulation is the value to detect a difference of 10 (5 – (– 5)) when the pattern of means is a linear growth pattern. Note that the power depends on both the pattern and the detectable difference.

Possible choices are displayed next using examples that assume that the factor has five categories.

- **List of Means**

  Enter a list of means directly into the List of Means box to the right.

- **Linear Up or Down**

  Up: -.50 -.25 0.0 .25 0.5. Down: .50 .25 0.0 -.25 -.50.

- **First Effect High or Low**

  High: .5 -.5 -.5 -.5 -.5. Low: -.5 .5 .5 .5 .5.

- **Last Effect High or Low**

  High: -.5 -.5 -.5 -.5 .5. Low: .5 .5 .5 .5 -.5.

- **First Half High or Low**

  High: .5 .5 .5 -.5 -.5. Low: -.5 -.5 -.5 .5 .5.

- **Zig Zag High or Low**

  High: .5 -.5 .5 -.5 .5. Low: -.5 .5 -.5 .5 -.5.

- **Interaction Effect Patterns**

  The effect patterns of interactions are formed as the Kronecker product of the individual factor patterns.

Detectable Difference

This is the difference between the largest and smallest effects associated with this factor or interaction. This represents the minimum detectable difference between any two levels (or factor-level combinations). The actual means used in the simulation are found by multiplying the Effects Pattern by this value.

Each value specified here results in a separate simulation. Note that, because of report and plot labeling, it is best if you only put multiple values in one of these boxes at a time. Otherwise, the plots will be labeled using the less-informative term ‘Combination’.

Example

Suppose that you selected Linear Up as the Effects Pattern of a five-level factor and set the detectable difference to 8. Further suppose that the Baseline Mean (below) is set to 20. The resulting means for this factor would be: 16 18 20 22 24.

Note that the maximum difference between any two of these means is 8.

List of Means

You can specify a list of means directly instead of specifying the pattern and detectable difference. When you do so, the effects pattern and detectable difference are calculated from the means you specify.

When you specify too many values for the number of levels in the factor, the extra values are ignored. When you specify too few values, the last value you specify is copied forward.

Using Columns in the Spreadsheet

You can specify the list of means in columns of the spreadsheet. When you do this, you enter the column(s) name(s) here using the equal sign.

For example, if you stored two sets of means, one in C5 and the other in C6, you would enter ‘=C5 C6’ here (or you could select the columns by pressing the button to the right). Once you have entered values into a spreadsheet, it is up to you to load that spreadsheet each time you run the simulation.

Note that a separate simulation is run for each column you specify.

Effect Size – Factors with Multiple Levels Within a Subject (Within)

These options specify the effect sizes of the within-subject factors (C and D).

Levels

Specify the number of levels (categories) in this factor. Typical values are from 2 to 8. Set this to a blank (or 0) to ignore the factor in the design.

Note that the number of time points is calculated as the product of the number of levels factors C and D.

The rest of these options behave as described in the Between-Subject section above.

Possible choices are displayed next using examples that assume that the factor has five categories.

Effect Size – Interactions

These options specify the effect sizes of any interactions that are specified.

Interaction Check Box

Check this box to include the corresponding interaction in the model. In most situations, you should include all appropriate interactions since these have an impact on the degrees of freedom of the F-tests. You can set their Detectable Difference near zero if you want to ignore them.

Occasionally, you will want to limit the number of interactions that you include in the model.

Note that the model is forced to be hierarchical. This means that if you include the three-way interaction ABC in the model, you must also include A, B, C, AB, AC, and BC.

Detectable Difference

This is the difference between the largest and smallest effects associated with this term. This represents the minimum detectable difference between any two factor-level combinations. The actual means used in the simulation are found by multiplying the effects pattern by this value. The effect pattern is created by forming the Kronecker product of the effect patterns of each factor in the interaction. The result is standardized so that the maximum difference in the pattern is 1.0. Each item in the pattern is multiplied by the Detectable Difference. The cell means are found by adding the appropriate effects for the main effects, interaction effects, and baseline mean.

Each value specified here results in a separate simulation. Note that, because of report and plot labeling, it is best if you only put multiple values in one of these boxes at a time. Otherwise, the plots will be labeled using the less-informative term ‘Combination’.

Covariance Tab

This tab specifies the variance-covariance matrix of the data that are generated during the simulation. Note that this structure does not have to match the structure of the mixed model that is fitted.

There are two components that can be specified: the random component and the residual component. If you specify a random component, you should set the residual component to 'Variance Pattern' (diagonal).

Random Component

Specify Using

Specify whether to include a Random Effects (subjects) component in the model. Possible choices are

- **None**

  Do not include a Random Component in the model.

- **Random Effects (Subjects)**

  Add a random effects component to the model. This component accounts for subject-to-subject variability. In this case, the Residual Component should be set to Variance Pattern (which does not include autocorrelations).

Random Component – Random Effects (Subjects) Parameter

Subject Variance

This is the value of random component's variance component. It is usually obtained from a previous run of a random effects mixed model. Since it is a variance, it must be positive.

When a Group Factor is used, a separate value must be supplied for each of the groups.

Residual Component

Specify Using

Specify how you want to specify the Residual Component.

- **Variance Pattern**

  This option indicates that the residual component includes only diagonal parameters (no autocorrelation).

- **Variance and Autocorrelation Patterns**

  This option indicates that the residual component includes off-diagonal parameters. Diagonal parameters represent variances and off-diagonal parameters represent autocorrelations.

- **Covariance Matrix in Spreadsheet**

  A covariance matrix is to be read in from the spreadsheet.

Residual Component – Variance (Diagonal) Pattern

Specify Variances

This option specifies how the variances of the residual component are specified. Possible choices are

- **Constant in V1**

  All variances are set to V1. All autocorrelations are set to zero.

  When a Grouping Factor is used, a list of variances can be read in from the spreadsheet. A column containing the group variances is specified with an equals sign, e.g. '=C1'. When this option is used, each row of the spreadsheet provides the variance for the corresponding group.

- **V1 to V2**

  The variances range from the V1 to V2 according to the Time Values. V1 is used for the first time value and V2 is used for the last time value.

  When a Grouping Factor is used, a list of V2 values can be read in from the spreadsheet. A column containing the V2's is specified with an equals sign, e.g. '=C2'. When this option is used, each row of the spreadsheet provides the variance for the corresponding group.

- **Variance List**

  A list of variances, one per time point, is specified here. The list(s) can also be specified as columns of the spreadsheet using the equals sign, e.g. '=C3'.

V1

The value of V1 is specified here.

When a Grouping Factor is used, a list of variances can be read in from the spreadsheet. A column containing the group variances is specified with an equals sign, e.g. '=C1'. When this option is used, each row of the spreadsheet provides the variance for the corresponding group.

V2

The value of V2 is specified here. The variances range from V1 to V2 proportional to the Time Values.

When a Grouping Factor is used, a list of variances can be read in from the spreadsheet. A column containing the group variances is specified with an equals sign, e.g. '=C2'. When this option is used, each row of the spreadsheet provides the variance for the corresponding group.

Variance List

A list of variances, one per time point, is specified here. The list can also be specified as a column of the spreadsheet using the equals sign, e.g. '=C3'.

Residual Component – Autocorrelation (Off-Diagonal) Pattern

Specify Autocorrelations

This option specifies the pattern of the autocorrelations in the Residual Component of the variance-covariance matrix. Three options are possible:

- **Constant in AC**

  The value of AC is used as the constant autocorrelation.

- **AR(1) in AC**

  The value of AC is used to generate a first order autocorrelation pattern. This pattern reduces the magnitude of the autocorrelation at each successive step by multiplying the value at the previous step by AC. Thus the pattern is AC AC*AC AC*AC*AC etc.

- **Autocorrelation List**

  A list of AC values, separated by blanks or commas, is used to specify the autocorrelation pattern across the time.

AC (Autocorrelation)

This is the autocorrelation, AC, between two measurements made on a subject at two time points that differ by one time unit. A value near 0 indicates low correlation. A value near 1 indicates high correlation. Its use depends on the selection in the Specify Autocorr's option.

Possible values range from -1 to 1. However, in this situation, a positive value is usually assumed, so the more realistic range is 0 to 1.

If Time Greater Than

This is the maximum time difference between two measurement points before the autocorrelation is set to a second autocorrelation value.

For example, you might wish to specify a constant AC of 0.4 for the first two time periods, and then, for any two measures that are greater than two time values apart, switch the autocorrelation to zero. If you have specified a Constant autocorrelation pattern, and set this value to ‘2’, the resulting autocorrelation pattern of the Residual Component would be

```
1.0  0.4  0.4  0.0  0.0
0.4  1.0  0.4  0.4  0.0
0.4  0.4  1.0  0.4  0.4
0.0  0.4  0.4  1.0  0.4
0.0  0.0  0.4  0.4  1.0
```

When you want to ignore this value, set it to a large number such as 100.

The cutoff value refers to the Time Values that you have specified elsewhere on this screen.

Then AC Becomes

This is the second autocorrelation value. It is used when the time difference between two measurement points is larger than the value to the left.

For example, suppose AC is 0.4, this value is 0.0, and ‘If Time Greater Than’ value is ‘2’. The resulting autocorrelation pattern of the Residual Component would be

1.0 0.4 0.4 0.0 0.0

0.4 1.0 0.4 0.4 0.0

0.4 0.4 1.0 0.4 0.4

0.0 0.4 0.4 1.0 0.4

0.0 0.0 0.4 0.4 1.0

When you want to ignore this value, set the ‘If Time Greater Than’ value to a large number such as 100.

RANGE: Since we usually assume a positive autocorrelation, this value ranges between 0 and 1.

Autocorrelation List

This is a list of autocorrelations, one for each time point. The number of autocorrelations must match the number of time values which is equal to the product of the number of levels for factors C and D.

Since this is a type of correlation, possible values range from -1 to 1. However, positive values are usually assumed, so the realistic range is 0 to 1. A value near 0 indicates low correlation. A value near 1 indicates high correlation.

You can alternatively enter the list of autocorrelations in a column of the spreadsheet and specify that column here. When the program finds an equals sign in this box, it reads in the values of the designated column. Blanks are converted to zeros. For example, if you have entered the autocorrelation list in column one, you would enter ‘=C1’ here.

If there are not enough values entered, the last value is copied forward.

Residual Component – Covariance Matrix in Spreadsheet

Covariance Matrix in Spreadsheet

This option designates the columns on the current spreadsheet holding the covariance matrix. The number of columns and number of rows with data must match the number of time periods at which the subjects are measured. The matrix must be positive definite.

Press the button at the right to select the columns from the spreadsheet.

Note that it is your responsibility to be certain that the spreadsheet is loaded with the correct values.

Miscellaneous Options

Grouping Factor

Specify a grouping factor--either Factor A or B.

When selected, a set of variance and autocorrelation parameters must be included for each level of this factor.

Time Values

This option specifies the time points at which measurements of the subjects are made. Often, measurements are made at equidistant points through time. But this is not always the case. The number of time points is the product of the number of levels of all within factors. The time metric influences the values of the variances as well as the correlations between two measurements on the same individual.

If the correlation matrix uses the time metric, the individual correlations are based on the formula: Corr(Yi,Yj)=AC^(|ti-tj|), where AC is the autocorrelation and ti and tj are two points in the time metric list. This formula allows you to easily specify many different types of correlation structures.

The syntax for entering the time metric is given next.

STEP <Start> <Inc>

Measurements are made at time intervals of length INC, beginning at START. For example, 'STEP 0 2' would generate the time series: 0, 2, 4, 6, etc.

RANGE <Min> <Max>

A set of equal-spaced time points is generated from the MIN value to the MAX value. This setting is very useful when you want to study the impact of increasing/decreasing the number of measurements per subject during the same period of time. That is, if the study will last five weeks, will the power of the statistical tests increase if you take ten measurements rather than five?

List

You may enter a list of values separated by blanks or commas. For example, you could enter 0 0.143 1 2 3 if times were 0 weeks, 1/7 week (day 1), 1 week, 2 weeks, 3 weeks.

Example

Suppose the number of time points is 6. Entering RANGE 0,10 will generate the Time Values: 0, 2, 4, 6, 8, 10. If the number of times is changed to 11, the Time Values will become: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Fitted Model Tab

This tab controls the variance-covariance matrix model that is actual fit during the solution of each simulation sample.

Fitted Variance-Covariance Model – Random Effects Component

Random Model

Specify whether to include a Random Effects (Subjects) component in the fitted model. Possible choices are

- **None**

  Do not include a Random Component in the fitted model.

- **Random Effects (Subjects)**

  Add a random subjects component to the fitted model. In this case, the Residual Component should be set to ‘Diagonal’ (which does not include autocorrelations).

Groups

Specify a grouping factor: either factor A or B.

When selected, a set of variance and autocorrelation parameters must be included for each level of this factor.

Fitted Variance-Covariance Model – Residual Component

Pattern

Specify the type of R matrix (Residual Component Pattern) to be generated. The default type is the Diagonal matrix. When terms are specified in the Random Model, this option should be set to Diagonal. A brief summary of the various structures follows.

R = Autocorrelation

Ri = ith autocorrelation

Rij = autocorrelation between ith and jth time points

S^2 = Sigma^2

Si^2 = Sigma^2 for ith time point

- **Diagonal**

```
|1 0 0 0|
|0 1 0 0|
|0 0 1 0|*S^2
|0 0 0 1|
```

- **Compound Symmetry**

```
|1 R R R|
|R 1 R R|
|R R 1 R|*S^2
|R R R 1|
```

- **AR(1)**

```
|1   R   R^2 R^3|
|R   1   R   R^2|
|R^2 R   1   R  |*S^2
|R^3 R^2 R   1  |
```

- **Toeplitz(i)**

  e.g. Toeplitz(3) =

  ```
  |1  R1 R2 0 |
  |R1 1  R1 R2|
  |R2 R1 1  R1|*S^2
  |0  R2 R1 1 |
  ```

- **Banded(i)**

  e.g. Banded(2) =

  ```
  |1 R 0 0|
  |R 1 R 0|
  |0 R 1 R|*S^2
  |0 0 R 1|
  ```

- **Unstructured**

  ```
  |1.00 R12 R13 R14|
  |R12 1.00 R23 R24|
  |R13 R23 1.00 R34|*S^2
  |R14 R24 R34 1.00|
  ```

  Heterogeneous covariance structures allow for nonconstant values for S^2.

  e.g. Diagonal - Heterogeneous =

  ```
  |S1^2 0    0    0   |
  |0    S2^2 0    0   |
  |0    0    S3^2 0   |
  |0    0    0    S4^2|
  ```

Force Positive Covariances

When checked, this option forces all covariances in the Random and Repeated Components (off-diagonal elements of the R matrix) to be non-negative. When this option is not checked, covariances can be negative.

Usually, negative covariances are okay and should be allowed. However, some Residual Component patterns such as ‘Compound Symmetry’ assume that covariances (autocorrelations) are positive.

Solution Options

Likelihood Type

Specify the type of likelihood equation to be solved. The options are:

- **MLE**

  The Maximum Likelihood solution has become less popular.

- **REML (recommended)**

  The Restricted Maximum Likelihood solution is recommended. It is the default in other software programs (such as SAS).

Solution Method

Specify the method to be used to solve the likelihood equations. The options are:

- **Newton-Raphson**

  This is an implementation of the popular 'gradient search' procedure for maximizing the likelihood equations. Whenever possible, we recommend that you use this method.

- **Fisher-Scoring**

  This is an intermediate step in the Newton-Raphson procedure. However, when the Newton-Raphson fails to converge, you may want to stop with this procedure.

- **MIVQUE**

  This non-iterative method is used to provide starting values for the Newton-Raphson method. For large problems, you may want to investigate the model using this method since it is much faster.

- **Differential Evolution**

  This grid search technique will often find a solution when the other methods fail to converge. However, it is painfully slow--often requiring hours to converge--and so should only be used as a last resort.

Solution Options – Newton-Raphson / Fisher Scoring Options

Fisher Scoring Iter’s

This is the maximum number of Fisher Scoring iterations that occur in the maximum likelihood finding process. When Solution Method is set to Newton-Raphson, up to this number of Fisher Scoring iterations occur before beginning Newton-Raphson iterations. We recommend setting this to ‘1’.

Newton-Raphson Iter’s

This is the maximum number of Newton-Raphson iterations that occur in the maximum likelihood finding process. When Solution Method is set to Newton-Raphson, Fisher-scoring iterations occur before beginning Newton-Raphson iterations. We recommend setting this to ‘5’.

Lambda

Each parameter’s change is multiplied by this value at each iteration. Usually, this value can be set to one. However, it may be necessary to set this value to 0.5 to implement step-halving: a process that is necessary when the Newton-Raphson diverges.

Note: this parameter only used by the Fisher-Scoring and Newton-Raphson methods.

Convergence Criterion

This procedure uses relative Hessian convergence (or the Relative Offset Orthogonality Convergence Criterion) as described by Bates and Watts (1981).

Recommended: The default value, 1E-8, will be adequate for many problems. When the routine fails to converge, try increasing the value to 1E-6.

Zero (Rounding)

This cutoff value is used by the least-squares algorithm to lessen the influence of rounding error. Values lower than this are reset to zero. If unexpected results are obtained, try using a smaller value, such as 1E-32. Note that 1E-5 is an abbreviation for the number 0.00001.

Recommended: 1E-10 or 1E-12.

Range: 1E-3 to 1E-40.

Variance Zero

When an estimated variance component (diagonal element) is less than this value, the variance is assumed to be zero and all reporting is terminated since the algorithm has not converged properly.

To correct this problem, remove the corresponding term from the Random Factors Model or simplify the Repeated Variance Pattern. Since the parameter is zero, why would you want to keep it?

Recommended: 1E-6 or 1E-8.

Range: 1E-3 to 1E-40.

Correlation Zero

When an estimated correlation (off-diagonal element) is less than this value, the correlation is assumed to be zero and all reporting is terminated since the algorithm has not converged properly.

To correct this problem, remove the corresponding term from the Random Factors Model or simplify the Repeated Variance Pattern. Since the parameter is zero, why would you want to keep it?

Recommended: 1E-6 or 1E-8.

Range: 1E-3 to 1E-40.

Max Retries

Specify the maximum number of retries to occur. During the maximum likelihood search process, the search may lead to an impossible combination of variance-covariance parameters (as defined by a matrix of variance-covariance parameters that is not positive definite). When such a combination arises, the search algorithm will begin again. Max Retries is the maximum number of times the process will re-start to avoid such combinations.

We recommend setting this to ‘1’.

Solution Options – Differential Evolution Options

Crossover Rate

This value controls the amount of movement of the differential evolution algorithm toward the current best. Larger values accelerate movement toward the current best, but reduce the chance of locating the global maximum. Smaller values improve the chances of finding the global, rather than a local, solution, but increase the number of iterations until convergence.

RANGE: Usually, a value between .5 and 1.0 is used.

RECOMMENDED: 0.9.

Mutation Rate

This value sets the mutation rate of the search algorithm. This is the probability that a parameter is set to a random value within the parameter space. It keeps the algorithm from stalling on a local maximum.

RANGE: Values between 0 and 1 are allowed.

RECOMMENDED: 0.9 for random coefficients (complex) models or 0.5 for random effects (simple) models.

Max Iter’s

Specify the maximum number of differential evolution iterations used by the differential evolution algorithm. A value between 100 and 200 is usually adequate. For large datasets, i.e., number of rows greater than 1000, you may want to reduce this number.

Min Relative Change

This parameter controls the convergence of the likelihood maximizer. When the relative change in the likelihoods from one generation to the next is less than this amount, the algorithm concludes that it has converged. The relative change is |L(g+1) - L(g)| / L(g) where L(g) is absolute value of the likelihood at generation 'g'. Note that the algorithm also terminates if the Maximum Generations are reached or if the number of individuals that are replaced in a generation is zero. The value 0.00000000001 (ten zeros) seems to work well in practice. Set this value to zero to ignore this convergence criterion.

Solutions/Iter’n

This is the number of trial points (solution sets) that are used by the differential evolution algorithm during each iteration. In the terminology of differential evolution, this is the population size.

RECOMMENDED: A value between 15 and 25 is recommended. More points may dramatically increase the running time. Fewer points may not allow the algorithm to converge.

Example 1 – Determining Power for Given Sample Size

Researchers are planning a study of the impact of a drug on heart rate. They want to evaluate the differences in heart rate among two age groups: 20-55 and over 55. Their experimental protocol calls for a baseline heart rate measurement, followed by administration of a certain level of the drug, followed by an additional measurement 30 minutes later. They want to be able to detect a 10% difference in heart rate among the age groups. They want to detect 5% difference in heart rate within an individual across time. They decide the experiment should detect interaction effects of the same magnitude as the within-subject factor. From a heart rate of 93, a 10% reduction gives 84, a difference of 9. Similarly, a 5% reduction within a subject would result in a heart rate of 88, a difference of 5.

They plan to analyze the data using a random effects model, setting the significance level at 0.05. Similar studies have found an average heart rate of 93, a subject variance of 11, and a residual variance of 11.

They want to look at the power for group sizes of 3, 5, 7, and 9.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Mixed Models** procedure window by clicking on **Means**, then **Mixed Models**, then **Mixed Models**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Sample Size	
n (Subjects Per Group)	**3 5 7 9**
=n's	**checked**
Alpha	
Alpha	**0.05**
Simulations	
Simulations	**100**
Effect Size	
Specify Effects Using	**Effects Patterns, etc.**
Factors Separating Subjects into Groups (Between)	
Levels (A)	**2**
Effects Pattern (A)	**Linear Up**
Detectable Difference (A)	**9**
Factors with Multiple Levels Within a Subject (Within)	
Levels (A)	**2**
Effects Pattern (A)	**Linear Up**
Detectable Difference (A)	**5**

Data Tab (continued)

Interactions
AC ....**checked**
Baseline Mean **93**

Covariance Tab

Random Component
Specify Using ....**Random Effects (Subjects)**

Random Effects Parameter
Subject Variance **11**

Residual Component
Specify Using ....**Variance Pattern**

Variance (Diagonal) Pattern
Specify Variances ....**Constant in V1**
V1 ....**11**

Fitted Model Tab

Fitted Variance-Covariance Model

Random Effects Component
Random Model ....**Random Effects (Subjects)**

Residual Component
Pattern...**Diagonal**

Solution Options
Likelihood Type ....**REML**
Solution Method ....**Newton-Raphson**

Reports Tab

Select Reports
To avoid accidental omission of output, all reports and plots are automatically displayed.

Report Options
Simulation Alpha **0.05**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Power Report for Each Design

```
Power by Design
                     Lower     Upper
                       95%       95%                 Minimum                Number of
Model                C.L. of   C.L. of            Detectable               Simulation      Factor
Term        Power    Power     Power   N(1)    N  Difference     Alpha       Samples      Pattern
A (2)      0.4300   0.3314    0.5329      3    6        9.00    0.0500           100    Linear Up
C (2)      0.4800   0.3790    0.5822      3    6        5.00    0.0500           100    Linear Up
AC         0.4900   0.3886    0.5920      3    6        5.00    0.0500           100

A (2)      0.8200   0.7305    0.8897      5   10        9.00    0.0500           100    Linear Up
C (2)      0.7900   0.6971    0.8651      5   10        5.00    0.0500           100    Linear Up
AC         0.9000   0.8238    0.9510      5   10        5.00    0.0500           100

A (2)      0.9600   0.9007    0.9890      7   14        9.00    0.0500           100    Linear Up
C (2)      0.8700   0.7880    0.9289      7   14        5.00    0.0500           100    Linear Up
AC         0.9500   0.8872    0.9836      7   14        5.00    0.0500           100

A (2)      0.9900   0.9455    0.9997      9   18        9.00    0.0500           100    Linear Up
C (2)      1.0000   0.9638    1.0000      9   18        5.00    0.0500           100    Linear Up
AC         0.9900   0.9455    0.9997      9   18        5.00    0.0500           100

Simulation Time: 65.78 seconds.
```

This report gives the simulated power for each term in the design for each value of N, one design at a time. It is useful when you want to compare the powers of the terms in the design at a specific sample size.

The definitions of each of the columns of the report are as follows.

Model Term

This column contains the identifying label of the term. The number of levels for a factor is given in parentheses.

Power

This is the simulated power for the term.

Lower and Upper 95% C.L. of Power

These are the lower and upper confidence limits of a 95% confidence interval for the simulated power. These are exact, distribution-free, confidence limits based on the binomial distribution.

N(1)

The value of the number of subjects per group.

N

The value of *N* is the total number of subjects in the study.

Minimum Detectable Difference

This is the value of the minimum detectable difference that was entered (or calculated) for this term. This is a measure of the magnitude of the effect size.

Alpha

Alpha is the significance level of the F-test for this term.

Number of Simulation Samples

This is the number of simulation samples for which the mixed model algorithm converged.

Factor Pattern

This is the type of factor pattern that was used with the detectable difference to form the group means.

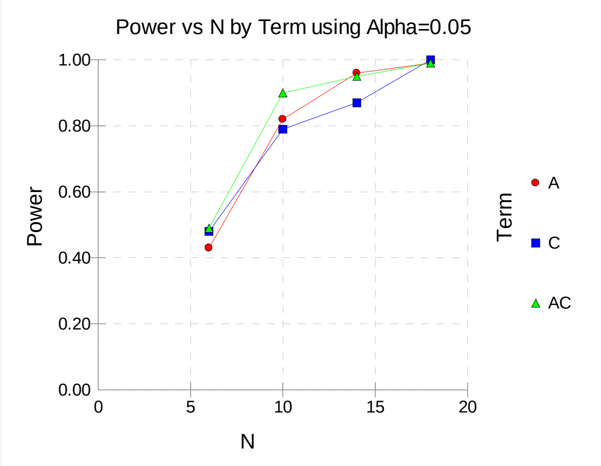
Term Reports

Power by Term

		Lower	Upper						
		95%	95%			Minimum		Number of	
Model		C.L. of	C.L. of			Detectable		Simulation	Factor
Term	Power	Power	Power	N(1)	N	Difference	Alpha	Samples	Pattern
A (2)	0.4300	0.3314	0.5329	3	6	9.00	0.0500	100	Linear Up
A (2)	0.8200	0.7305	0.8897	5	10	9.00	0.0500	100	Linear Up
A (2)	0.9600	0.9007	0.9890	7	14	9.00	0.0500	100	Linear Up
A (2)	0.9900	0.9455	0.9997	9	18	9.00	0.0500	100	Linear Up
C (2)	0.4800	0.3790	0.5822	3	6	5.00	0.0500	100	Linear Up
C (2)	0.7900	0.6971	0.8651	5	10	5.00	0.0500	100	Linear Up
C (2)	0.8700	0.7880	0.9289	7	14	5.00	0.0500	100	Linear Up
C (2)	1.0000	0.9638	1.0000	9	18	5.00	0.0500	100	Linear Up
AC	0.4900	0.3886	0.5920	3	6	5.00	0.0500	100	
AC	0.9000	0.8238	0.9510	5	10	5.00	0.0500	100	
AC	0.9500	0.8872	0.9836	7	14	5.00	0.0500	100	
AC	0.9900	0.9455	0.9997	9	18	5.00	0.0500	100	

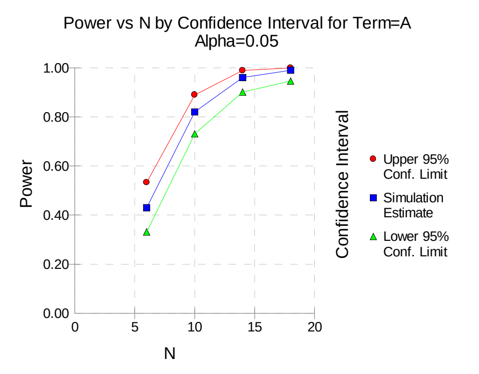
This report provides the same information as the previous Design report, so the item definitions are the same. It is sorted by term, to make comparison of powers across various sample sizes easier.

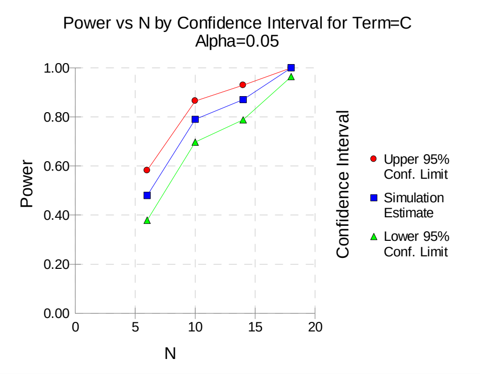
Whole Design Power Plots

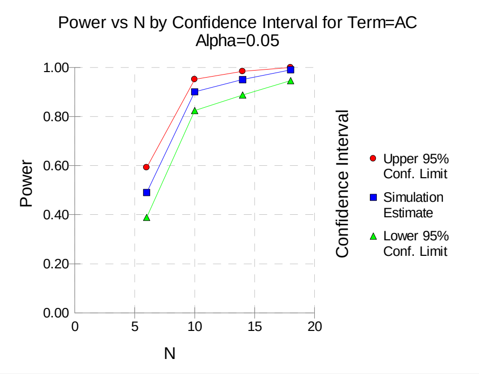


This plot shows the power values on the vertical axis and the sample sizes across the horizontal axis. The plot symbols represent different model terms.

Confidence Interval Power Plots

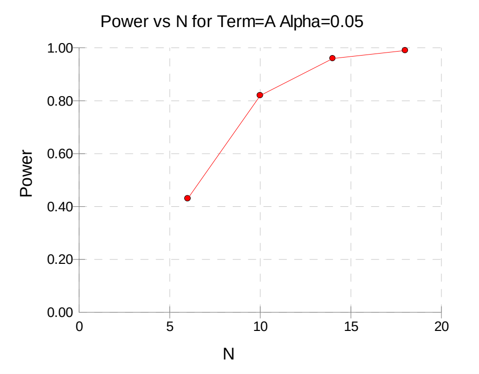


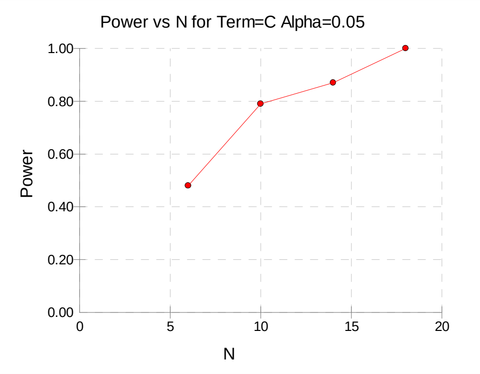


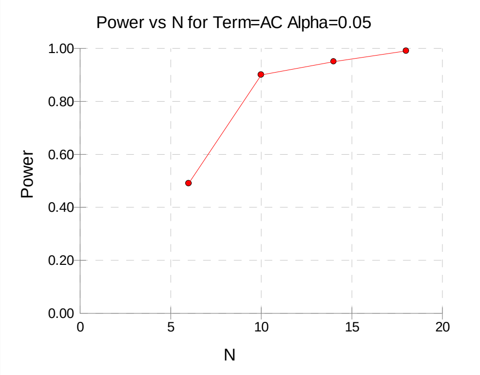


These plots show the confidence limits for the power for each term across the various sample sizes. Using these plots, you can quickly see for which sample sizes you might want to run a more precise (larger) simulation.

Individual Term Power Plots







These plots show the power for each term across the various sample sizes without including the confidence limits.

Detectable Differences for Each Term

```
Combination
No.                   A           C           AC
1                  9.00        5.00         5.00
```

This report shows the values of the Minimum Detectable Differences for each term. It is particularly useful to identify each combination when you enter several different values for various terms, so that there are several combinations.

Effect Pattern Section

```
---------- A ----------    ---------- C ----------    --------- AC ---------
Levels         Value       Levels         Value       Levels         Value
 1              -0.50       1              -0.50      1,1             0.50
 2               0.50       2               0.50      1,2            -0.50
                                                      2,1            -0.50
                                                      2,2             0.50
```

This report shows the patterns of each of the active terms in the model. These values are multiplied by the corresponding minimum detectable differences to form the effects associated with each term

Expanded Effect Pattern Section

```
---------- A ----------    ---------- C ----------    --------- AC ---------
Levels         Value       Levels         Value       Levels         Value
1,1             -0.50      1,1             -0.50      1,1             0.50
1,2             -0.50      1,2              0.50      1,2            -0.50
2,1              0.50      2,1             -0.50      2,1            -0.50
2,2              0.50      2,2              0.50      2,2             0.50
```

This report shows the effect patterns of each of the cells in the model. These values are multiplied by the corresponding minimum detectable difference and summed to form the cell means for each group in the design.

Hypothesized Means Matrix for Detectable Differences Combination 1

```
Factors          A1          A2
C1            88.50       92.50
C2            88.50      102.50
```

This report shows hypothesized means matrix for each combination of detectable differences. Note that the between factors are represented across the columns of the report and the within factors are represented down its rows.

Variance-Covariance Matrix

```
Factors          C1          C2
C1            22.00       11.00
C2            11.00       22.00
```

This report shows the variance-covariance matrix used to generate the data. You should check that they are correct.

Variance and Autocorrelations

```
Factors          C1          C2
C1            22.00        0.50
C2             0.50       22.00
```

This report shows the variances on the diagonal and the autocorrelations on the off-diagonal. You should check that they are correct.

Example 2 – Validation using Brown and Prescott

Brown and Prescott (2006) pages 268-269 present the following example of determining a sample size for a repeated measures design that is analyzed using a linear mixed model. Note that this analytic procedure is provided in PASS as the procedure entitled *Inequality Test for Two Means in a Repeated Measures Design*. That procedure is sometimes called the *time-averaged difference*.

In this example, a group size of 31 is found to achieve 80% power when the residual variance is 76, the autocorrelation is 0.53, the minimum detectable difference of a single between-subject factor is 5, the number of repeated measurements is 4, and the significance level is 0.05.

In this example, because of the long running time, the number of simulation samples will be set to 100. You can increase this to 500 or 1000 to obtain greater precision.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Mixed Models** procedure window by clicking on **Means**, then **Mixed Models**, then **Mixed Models**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Sample Size	
n (Subjects Per Group)	**31**
=n’s	**checked**
Alpha	
Alpha	**0.05**
Simulations	
Simulations	**100**
Effect Size	
Specify Effects Using	**Effects Patterns, etc.**
Factors Separating Subjects into Groups (Between)	
Levels (A)	**2**
Effects Pattern (A)	**Linear Up**
Detectable Difference (A)	**5**
Factors with Multiple Levels Within a Subject (Within)	
Levels (A)	**4**
Effects Pattern (A)	**Linear Up**
Detectable Difference (A)	**5**
Interactions	
AC	**unchecked**
Covariance Tab	
Random Component	
Specify Using	**none**
Residual Component	
Specify Using	**Variance and Autocorrelation Patterns**
Variance (Diagonal) Pattern	
Specify Variances	**Constant in V1**
V1	**76**
Autocorrelation Pattern	
Specify Autocorr’s	**Constant in AC**
AC	**0.53**
Fitted Model Tab	
Fitted Variance-Covariance Model	
Random Effects Component	
Random Model	**Random Effects (Subjects)**
Residual Component	
Pattern	**Diagonal**
Solution Options	
Likelihood Type	**REML**
Solution Method	**Newton-Raphson**

Output

Click the Run button to perform the calculations and generate the following output.

Power Report for Each Design

```
Power by Design
                         Lower     Upper
                           95%       95%                          Minimum              Number of
Model                   C.L. of   C.L. of                     Detectable             Simulation          Factor
Term          Power      Power     Power    N(1)      N    Difference     Alpha      Samples         Pattern
A (2)        0.7900     0.6971    0.8651      31     62          5.00    0.0500          100       Linear Up
C (4)        0.9800     0.9296    0.9976      31     62          5.00    0.0500          100       Linear Up

Simulation Time: 2.18 minutes.
```

Note that for just a 100 simulations, the power 0.79 has come out remarkably close to the actual value of 0.80 given by Brown and Prescott. The confidence interval, 0.70 to 0.87, is wide and could be run with a larger number of simulations to obtain a more precise answer.

Example 3 – Validation using Repeated Measures ANOVA

For balanced designs, LMM and Repeated Measures ANOVA can be used to analyze the same data. Although they both produce F-tests, the values of the F-test are different for the two tests.It is instructive to compare the power of these competing procedures.

In this example, we will determine the power for group sizes of 3, 9, and 15 when the random effects (subject) variance is 2, the residual variance is 2, the minimum detectable differences of all three terms is 1.0, the number of repeated measurements is 2, and the significance level is 0.05. A random effects model will be fit. The number of simulation samples will be set to 500.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Mixed Models** procedure window by clicking on **Means**, then **Mixed Models**, then **Mixed Models**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Sample Size	
n (Subjects Per Group)	**3 9 15**
=n's	**checked**
Alpha	
Alpha	**0.05**
Simulations	
Simulations	**500**
Data Tab (continued)	

Effect Size

Specify Effects Using .... **Effects Patterns, etc.**

Factors Separating Subjects into Groups (Between)

Levels (A) **2**

Effects Pattern (A).................................. **Linear Up**

Detectable Difference (A)........................ **1**

Factors with Multiple Levels Within a Subject (Within)

Levels (A) **2**

Effects Pattern (A).................................. **Linear Up**

Detectable Difference (A)........................ **1**

Interactions

AC .... **checked**

Detectable Difference (AC) **1**

Covariance Tab

Random Component

Specify Using .... **Random Effects (Subjects)**

Subject Variance **2**

Residual Component

Specify Using .... **Variance Pattern**

Variance (Diagonal) Pattern

Specify Variances .... **Constant in V1**

V1 .... **2**

Fitted Model Tab

Fitted Variance-Covariance Model

Random Effects Component

Random Model.. **Random Effects (Subjects)**

Residual Component

Pattern... **Diagonal**

Solution Options

Likelihood Type .... **REML**

Solution Method .... **Newton-Raphson**

Output

Click the Run button to perform the calculations and generate the following output.

Power Report for Each Design

```
Power by Design
                        Lower      Upper
                          95%        95%                        Minimum                  Number of
Model                  C.L. of    C.L. of                    Detectable              Simulation       Factor
Term        Power        Power      Power   N(1)       N     Difference      Alpha      Samples      Pattern
A (2)      0.0500       0.0326     0.0729      3       6           1.00     0.0500          500    Linear Up
C (2)      0.1440       0.1144     0.1779      3       6           1.00     0.0500          500    Linear Up
AC         0.1640       0.1326     0.1994      3       6           1.00     0.0500          500

A (2)      0.2220       0.1863     0.2610      9      18           1.00     0.0500          500    Linear Up
C (2)      0.4860       0.4414     0.5308      9      18           1.00     0.0500          500    Linear Up
AC         0.4920       0.4473     0.5368      9      18           1.00     0.0500          500

A (2)      0.3440       0.3024     0.3875     15      30           1.00     0.0500          500    Linear Up
C (2)      0.7540       0.7138     0.7912     15      30           1.00     0.0500          500    Linear Up
AC         0.7320       0.6909     0.7704     15      30           1.00     0.0500          500

Simulation Time: 5.54 minutes.
```

Next, we will run this same design through the Repeated Measures ANOVA procedure. To make it comparable, set both the between and within factors (B and W) to have two levels each with Sm set to 0.5. The covariance is set with SD1 equal to 2 and R1 equal to 0.5.

The results of this run are presented next.

Power Report for RMANOVA

```
Power Using RMANOVA
                                        Multiply      SD of     Standard
                                           Means    Effects    Deviation     Effect
Term    Test     Power       n       N        By       (Sm)      (Sigma)       Size      Alpha        Beta
B(2)    F       0.0859       3       6      1.00       0.50         1.73       0.29     0.0500      0.9141
W(2)    F       0.1588       3       6      1.00       0.50         1.00       0.50     0.0500      0.8412
BW      F       0.1588       3       6      1.00       0.50         1.00       0.50     0.0500      0.8412

B(2)    F       0.2105       9      18      1.00       0.50         1.73       0.29     0.0500      0.7895
W(2)    F       0.5134       9      18      1.00       0.50         1.00       0.50     0.0500      0.4866
BW      F       0.5134       9      18      1.00       0.50         1.00       0.50     0.0500      0.4866

B(2)    F       0.3328      15      30      1.00       0.50         1.73       0.29     0.0500      0.6672
W(2)    F       0.7529      15      30      1.00       0.50         1.00       0.50     0.0500      0.2471
BW      F       0.7529      15      30      1.00       0.50         1.00       0.50     0.0500      0.2471
```

The above results are summarized in the following table.

Comparing LMM and RMANOVA Power

		Lower	Upper		
		95%	95%		
Model		C.L. of	C.L. of	RMANOVA	
Term	Power	Power	Power	Power	N(1)
A (2)	0.0500	0.0326	0.0729	0.0859	3
C (2)	0.1440	0.1144	0.1779	0.1588	3
AC	0.1640	0.1326	0.1994	0.1588	3
A (2)	0.2220	0.1863	0.2610	0.2105	9
C (2)	0.4860	0.4414	0.5308	0.5134	9
AC	0.4920	0.4473	0.5368	0.5134	9
A (2)	0.3440	0.3024	0.3875	0.3328	15
C (2)	0.7540	0.7138	0.7912	0.7529	15
AC	0.7320	0.6909	0.7704	0.7529	15

By studying the values in this table, we see that two procedures agree very well for group sample sizes of 9 and 15. However, when the group sample size is 3, the power of RMANOVA is higher.

Example 4 – Heterogeneous Variances

One of the features offered by LMM is the ability to specify unequal group variances. This example will look at how to do this.

In this example, we will determine the power for group sizes of 6 when the random effects (subject) variance is 2 when A =1 and 4 with A=2, the residual variance is 2 when A=1 and 4 when A = 2, the minimum detectable differences of all three terms is 1.0, the number of repeated measurements is 2, and the significance level is 0.05. A random effects model will be fit. The number of simulation samples will be set to 100.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Mixed Models** procedure window by clicking on **Means**, then **Mixed Models**, then **Mixed Models**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Sample Size	
n (Subjects Per Group)	**6**
=n's	**checked**
Alpha	
Alpha	**0.05**
Simulations	
Simulations	**100**
Effect Size	
Specify Effects Using	**Effects Patterns, etc.**

Data Tab (continued)

Factors Separating Subjects into Groups (Between)

Levels (A) **2**

Effects Pattern (A).................................. **Linear Up**

Detectable Difference (A)....................... **1**

Factors with Multiple Levels Within a Subject (Within)

Levels (A) **2**

Effects Pattern (A).................................. **Linear Up**

Detectable Difference (A)....................... **1**

Interactions

AC .... **checked**

Detectable Difference (AC) **1**

Covariance Tab

Random Component

Specify Using .... **Random Effects (Subjects)**

Subject Variance **2 4**

Residual Component

Specify Using .... **Variance Pattern**

Variance (Diagonal) Pattern

Specify Variances .... **Constant in V1**

V1 .... **2 4**

Miscellaneous Options

Grouping Factor .... **A**

Fitted Model Tab

Fitted Variance-Covariance Model

Random Effects Component

Random Model....................................... **Random Effects (Subjects)**

Groups .... **A**

Residual Component

Pattern.. **Diagonal**

Groups .... **A**

Solution Options

Likelihood Type...................................... **REML**

Solution Method .... **Newton-Raphson**

Output

Click the Run button to perform the calculations and generate the following output.

Power Report for Each Design

```
Power by Design
                      Lower      Upper
                        95%        95%                      Minimum                  Number of
Model                C.L. of    C.L. of                  Detectable               Simulation          Factor
Term       Power       Power      Power   N(1)       N   Difference      Alpha       Samples          Pattern
A (2)     0.1500      0.0865     0.2353      6      12         1.00     0.0500           100        Linear Up
C (2)     0.2200      0.1433     0.3139      6      12         1.00     0.0500           100        Linear Up
AC        0.2800      0.1948     0.3787      6      12         1.00     0.0500           100

Variance-Covariance Matrix
Factors             C1           C2
A1C1              4.00         2.00     Note that now there are two variance-covariance matrices displayed.
A1C2              2.00         4.00
A2C1              8.00         4.00
A2C2              4.00         8.00

Variances and Autocorrelations
Factors             C1           C2
A1C1              4.00         0.50
A1C2              0.50         4.00
A2C1              8.00         0.50
A2C2              0.50         8.00
```

Notice how easy it is to obtain a power analysis for the case of unequal variances.

Chapter 575

Multiple Comparisons

Introduction

This module computes sample sizes for multiple comparison procedures. The term "*multiple comparison*" refers to the individual comparison of two means selected from a larger set of means. The module emphasizes one-way analysis of variance designs that use one of three multiple-comparison methods: Tukey's all pairs (MCA), comparisons with the best (MCB), or Dunnett's all versus a control (MCC). Because these sample sizes may be substantially different from those required for the usual *F* test, a separate module is provided to compute them.

There are only a few articles in the statistical literature on the computation of sample sizes for multiple comparison designs. This module is based almost entirely on the book by Hsu (1996). We can give only a brief outline of the subject here. Users who want more details are referred to Hsu's book.

Although this module is capable of computing sample sizes for unbalanced designs, it emphasizes balanced designs.

Technical Details

The One-Way Analysis of Variance Design

The summarized discussion that follows is based on the common, one-way analysis of variance design. Suppose the responses Y_{ij} in *k* groups each follow a normal distribution with respective means, $\mu_1, \mu_2, \cdots, \mu_k$, and unknown variance, σ^2. Let $n_1, n_2, \cdots, n_k$ denote the number of subjects in each group.

The analysis of these responses is based on the sample means

$$\hat{\mu}_i = \overline{Y}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

and the pooled sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2}{\sum_{i=1}^{k}\left(n_i - 1\right)}$$

The *F* test is the usual method for analyzing such a design, and tests whether all of the means are equal. However, a significant *F* test does not indicate which of the groups are different, only that at least one is different. The analyst is left with the problem of determining which group(s) is(are) different and by how much.

The probability statement associated with the *F* test is simple, straightforward, and easy to interpret. However, when several simultaneous comparisons are made among the group means, the interpretation of individual probability statements becomes much more complex. This is called the problem of *multiplicity*. *Multiplicity* here refers to the fact that the probability of making at least one incorrect decision increases as the number of statistical tests increases. The method of *multiple comparisons* has been developed to account for such multiplicity.

Power Calculations for Multiple Comparisons

For technical reasons, the definition of power in the case of multiple comparisons is different from the usual definition. Following Hsu (1996) page 237, power is defined as follows.

Using a $1-\alpha$ simultaneous confidence interval multiple comparison method, power is the probability that the confidence intervals cover the true parameter values and are sufficiently narrow. Power is still defined to be $1-\beta$. Note that $1-\beta < 1-\alpha$. Here, *narrow* refers to the width of the confidence intervals. The definition says that the confidence intervals should be as narrow as possible while still including the true parameter values. This definition may be restated as the probability that the simultaneous confidence intervals are *correct* and *useful*.

The parameter ω represents the maximum width of any of the individual confidence intervals in the set of simultaneous confidence intervals. Thus, ω is used to specify the narrowness of the confidence intervals.

Multiple Comparisons with a Control (MCC)

A common experimental design compares one or more *treatment* groups with a *control* group. The control group may receive a placebo, the standard treatment, or even an experimental treatment. The distinguishing feature is that the mean response of each of the other groups is to be compared with this control group.

We arbitrarily assume that the last group (group *k*), is the control group. The *k*-1 parameters of primary interest are

$$\begin{gathered}
\delta_1 = \mu_1 - \mu_k \\
\delta_2 = \mu_2 - \mu_k \\
\vdots \\
\delta_i = \mu_i - \mu_k \\
\vdots \\
\delta_{k-1} = \mu_{k-1} - \mu_k
\end{gathered}$$

In this situation, Dunnett's method provides simultaneous, one- or two-sided confidence intervals for all of these parameters.

The one-sided confidence intervals, $\delta_1, \cdots, \delta_{k-1}$, are specified as follows:

$$\Pr\left(\delta_i > \hat{\mu}_i - \hat{\mu}_k - q_{\alpha, df, \lambda_1, \cdots, \lambda_{k-1}} \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_k}} \;\; for\ i = 1, \cdots, k-1\right) = 1 - \alpha$$

where

$$\lambda_i = \sqrt{\frac{n_i}{n_i + n_k}}$$

and q, found by numerical integration, is the solution to

$$\int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^{k-1} \left[\Phi\left(\frac{\lambda_i z + qs}{\sqrt{1 - \lambda^2}}\right)\right] d\Phi(z) \gamma(s) ds = 1 - \alpha$$

where $\Phi(z)$ is the standard normal distribution function and $\gamma(z)$ is the density of $\hat{\sigma} / \sigma$.

The two-sided confidence intervals for $\delta_1, \cdots, \delta_{k-1}$ are specified as follows:

$$\Pr\left(\delta_i \in \hat{\mu}_i - \hat{\mu}_k - \left|q_{\alpha, df, \lambda_1, \cdots, \lambda_{k-1}}\right| \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_k}} \;\; for\ i = 1, \cdots, k-1\right) = 1 - \alpha$$

where

$$\lambda_i = \sqrt{\frac{n_i}{n_i + n_k}}$$

and $|q|$, found by numerical integration, is the solution to

$$\int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^{k-1} \left[\Phi\left(\frac{\lambda_i z + |q| s}{\sqrt{1 - \lambda^2}}\right) - \Phi\left(\frac{\lambda_i z - |q| s}{\sqrt{1 - \lambda^2}}\right)\right] d\Phi(z) \gamma(s) ds = 1 - \alpha$$

where $\Phi(z)$ is the standard normal distribution function and $\gamma(z)$ is the density of $\hat{\sigma} / \sigma$.

Interpretation of Dunnett's Simultaneous Confidence Intervals

There is a specific interpretation given for Dunnett's method. It provides a set of confidence intervals calculated so that, if the normality and equal-variance assumptions are valid, the probability that all of the *k*-1 confidence intervals enclose the true values of $\delta_1, \cdots, \delta_{k-1}$ is $1 - \alpha$. The presentation below is for the two-sided case. The one-sided case is the same as for the MCB case.

Sample Size and Power – Balanced Case

Using the modified definition of power, the two-sided case is outlined as follows.

$$\begin{aligned}
&\Pr\left[\left(\text{simultaneous coverage}\right) \text{ and } \left(\text{narrow}\right)\right] \\
&= \Pr\left[\left(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm |q|\hat{\sigma}\sqrt{2/n} \text{ for } i = 1, \cdots, k-1\right) \text{ and } \left(|q|\hat{\sigma}\sqrt{2/n} < \omega/2\right)\right] \\
&= k\int_0^u \int_{-\infty}^{\infty} \left[\Phi\left(z + \sqrt{2}|q|s\right) - \Phi\left(z - \sqrt{2}|q|s\right)\right]^{k-1} d\Phi(z)\gamma(s)ds \\
&\geq 1 - \beta
\end{aligned}$$

where

$$u = \frac{\omega/2}{\sigma|q|\sqrt{\frac{2}{n}}}$$

This calculation is made using the algorithm developed by Hsu (1996).

To reiterate, this calculation requires you to specify the minimum value of $\omega = \mu_i - \mu_j$ that you want to detect, the group sample size, n, the power, $1 - \beta$, the significance level, α, and the within-group standard deviation, σ.

Sample Size and Power – Unbalanced Case

Using the modified definition of power, the unbalanced case is outlined as follows.

$$\begin{aligned}
&\Pr\left[(\text{simultaneous coverage}) \text{ and } (\text{narrow})\right] \\
&= \Pr\left[\left(\mu_i - \mu_k \in \hat{\mu}_i - \hat{\mu}_k \pm |q|\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_k}} \text{ for } i = 1, \ldots, k-1\right)\right. \\
&\qquad \left. \text{and } \left(\min_{i<k}\left[|q|\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}\right]\right) < \omega/2\right] \\
&= k\int_0^u \int_{-\infty}^{\infty} \left[\Phi(z) - \Phi\left(z - \sqrt{2}|q|s\right)\right]^{k-1} d\Phi(z)\gamma(s)ds \\
&\geq 1 - \beta
\end{aligned}$$

where

$$u = \frac{\omega/2}{\min_{i<k}\left[\sigma|q|\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}\right]}$$

This calculation is made using the algorithm developed by Hsu (1996).

To reiterate, this calculation requires you to specify the minimum value of $\omega = \mu_i - \mu_j$ that you want to detect, the group sample sizes, $n_1, n_2, \cdots, n_k$, the power, $1 - \beta$, the significance level, α, and the within-group standard deviation, σ.

Multiple Comparisons with the Best (MCB)

The method of multiple comparisons with the best (champion) is used in situations in which the best group (we will assume the best is the largest, but it could just as well be the smallest) is desired. Because of sampling variation, the group with the largest sample mean may not actually be the group with the largest population mean. The following methodology has been developed to analyze data in this situation.

Perhaps the most obvious way to define the parameters in this situation is as follows

$$\max_{j=1,\cdots,k} \mu_j - \mu_i,\ for\ i = 1,\cdots,k$$

Obviously, the group for which all of these values are positive will correspond to the group with the largest mean.

Another way of looking at this, which has some advantages, is to use the parameters

$$\theta_i = \mu_i - \max_{i \neq j} \mu_j,\ for\ i = 1,\cdots,k$$

since, if $\theta_i > 0$, group *i* is the best.

Hsu (1996) recommends using constrained MCB inference in which the intervals are constrained to include zero. Hsu recommends this because inferences about which group is best are sharper. For example, a confidence interval for θ_i whose lower limit is 0 indicates that group *i* is the best. Similarly, a confidence interval for θ_i whose upper limit is 0 indicates that group *i* is not the best.

Hsu (1996) shows that $100(1-\alpha)\%$ simultaneous confidence intervals for θ_i are given by

$$-\min\left[0,\left(\hat{\theta}_i - q^i_{\alpha,df,\lambda_1,\cdots,\lambda_{k-1}}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_k}}\right)\right], \max\left[0,\left(\hat{\theta}_i + q^i_{\alpha,df,\lambda_1,\cdots,\lambda_{k-1}}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_k}}\right)\right], i = 1,\cdots,k$$

where q^i is found using Dunnett’s one-sided procedure discussed above assuming that group *i* is the control group.

Sample Size and Power – Balanced Case

Using the modified definition of power, the balanced case is outlined as follows

$$\begin{aligned}
&\Pr\left[(\text{simultaneous coverage})\text{ and }(\text{narrow})\right] \\
&= \Pr\left[\begin{array}{l}\left(\begin{array}{l}-\min\left(0, \hat{\mu}_i - \max\limits_{j\neq i}\left(\hat{\mu}_j\right) - q\hat{\sigma}\sqrt{2/n}\right) \leq \\ \mu_i - \max\limits_{j\neq i}\left(\mu_j\right) \leq \\ \max\left(0, \hat{\mu}_i - \max\limits_{j\neq i}\left(\hat{\mu}_j\right) + q\hat{\sigma}\sqrt{2/n}\right)\ \text{for } i = 1,\ldots,k\end{array}\right) \\ \text{and } \left(q\hat{\sigma}\sqrt{2/n} < \omega/2\right)\end{array}\right] \\
&= k\int_0^u \int_{-\infty}^{\infty}\left[\Phi\left(z+\sqrt{2}qs\right)\right]^{k-1} d\Phi(z)\gamma(s)ds \\
&\geq 1-\beta
\end{aligned}$$

where

$$u = \frac{\omega / 2}{\sigma q \sqrt{\frac{2}{n}}}$$

This calculation is made using the algorithm developed by Hsu (1996).

To reiterate, this calculation requires you to specify the minimum value of $\omega = \mu_i - \mu_j$ that you want to detect, the group sample size, *n*, the power, $1 - \beta$, the significance level, α, and the within-group standard deviation, σ.

Sample Size and Power – Unbalanced Case

Using the modified definition of power, the unbalanced case is outlined as follows

$$\begin{aligned}
&\Pr\Big[\big(\text{simultaneous coverage}\big) \text{ and } \big(\text{narrow}\big)\Big] \\
&= \Pr\left[\begin{array}{l}
\left(\begin{array}{l}
-\min\left(0, \hat{\mu}_i - \max\limits_{j \neq i}\left(\hat{\mu}_j\right) - q^i \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right) \leq \\
\mu_i - \max\limits_{j \neq i}\left(\mu_j\right) \leq \\
\max\left(0, \hat{\mu}_i - \max\limits_{j \neq i}\left(\hat{\mu}_j\right) + q^i \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right) \text{ for } i = 1, \cdots, k
\end{array}\right) \\
\text{and } \left(\min\limits_{j \neq i}\left[q^i \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\right] < \omega / 2\right)
\end{array}\right] \\
&= k \int_0^u \int_{-\infty}^{\infty} \left[\Phi\left(z + \sqrt{2}|q|s\right)\right]^{k-1} d\Phi(z)\gamma(s)\,ds \\
&\geq 1 - \beta
\end{aligned}$$

where

$$u = \max_{i \neq j}\left(\frac{\omega / 2}{\sigma\left|q^i\right|\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}\right)$$

This calculation is made using the algorithm developed by Hsu (1996).

To reiterate, this calculation requires you to specify the minimum value of $\omega = \mu_i - \mu_j$ that you want to detect, the group sample sizes, $n_1, n_2, \cdots, n_k$, the power, $1 - \beta$, the significance level, α, and the within-group standard deviation, σ.

All-Pairwise Comparisons (MCA)

In this case you are interested in all possible pairwise comparisons of the group means. There are *k*(*k*-1)/2 such comparisons. A popular method in this case is that developed by Tukey.

Balanced Case

The Tukey method provides simultaneous, two-sided confidence intervals. They are specified as follows

$$\Pr\left(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_k \pm |q^*| \hat{\sigma} \sqrt{\frac{2}{n}} \ \textit{for } i \neq j \right) = 1 - \alpha$$

When all the sample sizes are equal, q^* may be found by numerical integration as the solution to the equation

$$\int_0^{\infty} \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \left[\Phi(z) - \Phi\left(z - \sqrt{2}|q^*|s\right) \right] d\Phi(z) \gamma(s) ds = 1 - \alpha$$

where $\Phi(z)$ is the standard normal distribution function and $\gamma(z)$ is the density of $\hat{\sigma} / \sigma$. Note that $q' = \sqrt{2}|q^*|$ is the critical value of the Studentized range distribution.

Sample Size and Power

Using the modified definition of power, the balanced case is outlined as follows

$$\begin{aligned}
&\Pr\left[(\text{simultaneous coverage}) \text{ and } (\text{narrow}) \right] \\
&= \Pr\left[\left(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm |q^*| \hat{\sigma} \sqrt{2/n} \text{ for all } i \neq j \right) \text{ and } \left(|q^*| \hat{\sigma} \sqrt{2/n} < \omega/2 \right) \right] \\
&= k \int_0^{u} \int_{-\infty}^{\infty} \left[\Phi(z) - \Phi\left(z - \sqrt{2}|q^*|s\right) \right]^{k-1} d\Phi(z) \gamma(s) ds \\
&\geq 1 - \beta
\end{aligned}$$

where

$$u = \frac{\omega / 2}{\sigma |q^*| \sqrt{\frac{2}{n}}}$$

This calculation is made using the algorithm developed by Hsu (1996).

To reiterate, this calculation requires you to specify the minimum value of $\omega = \mu_i - \mu_j$ that you want to detect, the group sample size, *n*, the power, $1 - \beta$, the significance level, α, and the within-group standard deviation, σ.

Unbalanced Case

The simultaneous, two-sided confidence intervals are specified as follows

$$\Pr\left(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_k \pm \left|q^e\right|\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_k}}\ \ for\ i \neq j\right) = 1-\alpha$$

Unfortunately, $\left|q^e\right|$ cannot be calculated as a double integral as in previous cases. Instead, the Tukey-Kramer approximate solution is used. Their proposal is to use $\left|q^*\right|$ in place of $\left|q^e\right|$.

Sample Size and Power

Using the modified definition of power, the unbalanced case is outlined as follows

$$\begin{aligned}
&\Pr\left[(\text{simultaneous coverage})\text{ and }(\text{narrow})\right] \\
&= \Pr\left[\left(\mu_i - \mu_j \in \hat{\mu}_i - \hat{\mu}_j \pm |q^*|\,\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\text{ for all } i \neq j\right)\right. \\
&\qquad\qquad \left.\text{and}\left(\min_{i\neq j}\left[|q^*|\,\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right]\right) < \omega/2\right] \\
&= k\int_0^u \int_{-\infty}^{\infty}\left[\Phi(z) - \Phi\left(z - \sqrt{2}\,|q^*|\,s\right)\right]^{k-1} d\Phi(z)\gamma(s)ds \\
&\geq 1-\beta
\end{aligned}$$

where

$$u = \frac{\omega/2}{\min_{i\neq j}\left[\sigma\left|q^*\right|\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right]}$$

This calculation is made using the algorithm developed by Hsu (1996).

To reiterate, this calculation requires you to specify the minimum value of $\omega = \mu_i - \mu_j$ that you want to detect, the group sample sizes, $n_1, n_2, \cdots, n_k$, the power, $1-\beta$, the significance level, α, and the within-group standard deviation, σ.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for using the other parameters. Under most situations you will select either *Power and Beta* for a power analysis or *n* for sample size determination.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. In the case of multiple comparisons, power is the probability that the confidence intervals will cover the true parameter values and be sufficiently narrow to be useful. This is a modified definition of power; since beta equals 1-power, it is has a modified definition as well.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size / Groups – Sample Size Multiplier

n (Sample Size Multiplier)

This is the base, per-group sample size. One or more values separated by blanks or commas may be entered. A separate analysis is performed for each value listed here.

The group samples sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by *m1*, *m2*, *m3*, …, *mk* and this value is represented by *n*, the group sample sizes *N1*, *N2*, *N3*, ..., *Nk* are calculated as follows:

N1=[n(m1)]

N2=[n(m2)]

N3=[n(m3)]

etc.

where the operator, [*X*] means the next integer after *X*, e.g. [3.1]=4.

For example, suppose there are three groups and the Group Sample Size Pattern is set to *1,2,3*. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to *2,4,6,8,10*, five sets of group sample sizes will be generated and an analysis run for each. These sets are:

2	4	6
4	8	12
6	12	18
8	16	24
10	20	30

As a second example, suppose there are three groups and the Group Sample Size Pattern is *0.2,0.3,0.5*. When the fractional Pattern values sum to one, n can be interpreted as the total sample size of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

(0.2)12 = 2.4 which is rounded up to the next whole integer, 3.

(0.3)12 = 3.6 which is rounded up to the next whole integer, 4.

(0.5)12 = 6.

Note that in this case, 3+4+6 does not equal *n* (which is 12). This can happen because of rounding.

Sample Size / Groups – Groups

k (Number of Groups)

This is the number of group means being compared. It must be greater than or equal to three.

You can enter a list of values, in which case a separate analysis will be calculated for each value. Commas or blanks may separate the numbers. A TO-BY list may also be used.

Note that the number of items used in the Hypothesized Means box and the Group Sample Size Pattern box is controlled by this number.

Examples:

3,4,5

4 5 6

3 to 11 by 2

Group Sample Size Pattern

A set of positive, numeric values (one for each group) is entered here. The sample size of group *i* is found by multiplying the *i*th number from this list times the value of *n* and rounding up to the next whole number. The number of values must match the number of groups, *k*. When too few numbers are entered, 1's are added. When too many numbers are entered, the extras are ignored.

Note that in the case of Dunnett's test, the last group corresponds to the control group. Thus, if you want to study the implications of having a group with a different sample size for the control group, you should specify that sample size in the last position.

- **Equal**

  If all sample sizes are to be equal, enter "Equal" here and the desired sample size in n. A set of *k* 1's will be used. This will result in *N1* = *N2* = *N3* = *n*. That is, all sample sizes are equal to *n*.

Effect Size

Minimum Detectable Difference

Specify one or more values of the minimum detectable difference. This is the smallest difference between any two group means that is to be detectable by the experiment. Note that this is a positive amount. This value is set by the researcher to represent the smallest difference between two means that will be of practical significance to other researchers. Note that in the case of Dunnett's test, this is the maximum difference between each treatment and the control.

Examples:

3

3 4 5

10 to 20 by 2

S (Standard Deviation of Subjects)

This option refers to σ, the standard deviation within a group. It represents the variability from subject to subject that occurs when the subjects are treated identically. It is assumed to be the same for all groups. This value is approximated in an analysis of variance table by the square root of the mean square error.

Since they are positive square roots, the numbers must be strictly greater than zero. You can press the *SD* button to obtain further help on estimating the standard deviation.

Note that if you are using this procedure to test a factor (such as an interaction) from a more complex design, the value of standard deviation is estimated by the square root of the mean square of the term that is used as the denominator in the *F* test.

You can enter a list of values separated by blanks or commas, in which case a separate analysis will be calculated for each value.

Examples of valid entries:

1,4,7,10

1 4 7 10

1 to 10 by 3

Test

Type of Multiple Comparison

Specify the type of multiple comparison test to be analyzed. The tests available are

- **All Pairs - Tukey Kramer**
  All possible paired comparisons.

- **With Best - Hsu**
  Constrained comparisons with the best.

- **With Control - Dunnett**
  Two-sided versus a control group.

Iterations/Precision Tab

The Iterations/Precision tab contains parameters that control the progress and termination of the iterative procedures.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank.

Precision

Search Precision

Specify the precision to be used in the routines that search for the value of the minimum detectable difference. Note that the closer this value is to zero, the longer a search will take.

Example 1 – Calculating Power

An experiment is being designed to compare the means of four groups using the Tukey-Kramer pairwise multiple comparison test with a significance level of 0.05. Previous studies indicate that the standard deviation is 5.3. The typical mean response level is 63.4. The researcher believes that a 25% increase in the mean will be of interest to others. Since 0.25(63.4) = 15.85, this is the number that will be used as the minimum detectable difference.

To better understand the relationship between power and sample size, the researcher wants to compute the power for several group sample sizes between 2 and 14. The sample sizes will be equal across all groups.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Comparisons (Analytic)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**2 to 14 by 2**
k (Number of Groups)	**4**
Group Sample Size Pattern	**Equal**
Minimum Detectable Difference	**15.85**
S (Standard Deviation of Subjects)	**5.3**
Type of Multiple Comparison	**All Pairs - Tukey Kramer**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results Report

```
Numeric Results for Multiple Comparison Test: Tukey-Kramer (Pairwise)
          Average                                       Minimum     Standard
             Size              Total                 Detectable    Deviation
Power         (n)      k          N    Alpha     Beta  Difference          (S)        Diff / S
0.0113       2.00      4          8   0.0500   0.9887       15.85         5.30          2.9906
0.0666       4.00      4         16   0.0500   0.9334       15.85         5.30          2.9906
0.3171       6.00      4         24   0.0500   0.6829       15.85         5.30          2.9906
0.7371       8.00      4         32   0.0500   0.2629       15.85         5.30          2.9906
0.9301      10.00      4         40   0.0500   0.0699       15.85         5.30          2.9906
0.9497      12.00      4         48   0.0500   0.0503       15.85         5.30          2.9906
0.9500      14.00      4         56   0.0500   0.0500       15.85         5.30          2.9906
```

```
Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
n is the average group sample size.
k is the number of groups.
Total N is the total sample size of all groups.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
The Minimum Detectable Difference between any two group means.
S is the within group standard deviation.
The Diff / S is the ratio of Min. Detect. Diff. to standard deviation.

Summary Statements
In a single factor ANOVA study, sample sizes of 2, 2, 2, and 2 are obtained from the 4 groups
whose means are to be compared. The total sample of 8 subjects achieves 1% power to detect a
difference of at least 15.85 using the Tukey-Kramer (Pairwise) multiple comparison test at a
0.0500 significance level. The common standard deviation within a group is assumed to be 5.30.
```

This report shows the numeric results of this power study. Following are the definitions of the columns of the report.

Power

This is the probability that the confidence intervals will cover the true parameter values and be sufficiently narrow to be useful.

Average n

The average of the group sample sizes.

k

The number of groups.

Total N

The total sample size of the study.

Alpha

The probability of rejecting a true null hypothesis. This is often called the significance level.

Beta

Equal to 1- power. Note the definition of power above.

Minimum Detectable Difference

This is the value of the minimum detectable difference. This is the minimum difference between two means that is thought to be of practical importance. Note that in the case of Dunnett’s test, this is the minimum difference between a treatment mean and the control mean that is of practical importance.

Standard Deviation (S)

This is the within-group standard deviation. It was set in the Data window.

Diff / S

This is an index of relative difference between the means standardized by dividing by the standard deviation. This value can be used to make comparisons among studies.

Detailed Results Report

Tukey-Kramer Test Details

Group	n	Percent n of Total N	Alpha	Power	Minimum Detectable Difference	Standard Deviation
1	2	25.00	0.0500	0.0113	15.85	5.30
2	2	25.00				
3	2	25.00				
4	2	25.00				
Total	8	100.00				

This report shows the details of each row of the previous report.

Group

The group identification number is shown on each line. The second to the last line represents the last group. When Dunnett's test has been selected, this line represents the control group.

The last line, labeled *Total*, gives the total for all the groups.

n

This is the sample size of each group. This column is especially useful when the sample sizes are unequal.

Percent n of Total N

This is the percentage of the total sample that is allocated to each group.

Alpha

The probability of rejecting a true null hypothesis. This is often called the significance level.

Power

This is the probability that the confidence intervals will cover the true parameter values and be sufficiently narrow to be useful.

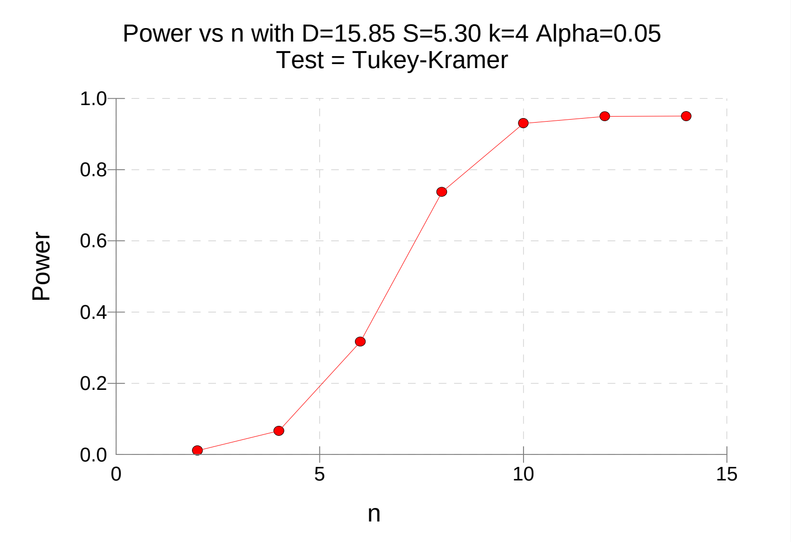
Minimum Detectable Difference

This is the value of the minimum detectable difference. This is the minimum difference between two means that is thought to be of practical importance. Note that in the case of Dunnett's test, this is the minimum difference between a treatment mean and the control mean that is of practical importance.

Standard Deviation (S)

This is the within-group standard deviation. It was set in the Data window.

Plots Section



This plot gives a visual presentation to the results in the Numeric Report. We can see the impact on the power of increasing the sample size.

When you create one of these plots, it is important to use trial and error to find an appropriate range for the horizontal variable so that you have results with both low and high power.

Example 2 – Power after Dunnett's Test

This example covers the situation in which you are calculating the power of Dunnett's test on data that have already been collected and analyzed.

An experiment included a control group and two treatment groups. Each group had seven individuals. A single response was measured for each individual and recorded in the following table.

Control	T1	T2
554	774	786
447	465	536
356	759	653
452	646	685
674	547	658
654	665	669

When analyzed using the one-way analysis of variance procedure in ***NCSS***, the following results were obtained.

```
Analysis of Variance Table

Source                          Sum of          Mean                   Prob
Term               DF          Squares        Square      F-Ratio      Level
A ( ... )           2          75629.8       37814.9         3.28      0.061
S(A)               18         207743.4       11541.3
Total (Adjusted)   20         283373.3
Total              21

Dunnett's Simultaneous Confidence Intervals for Treatment vs. Control
Response: Control,T1,T2
Term A:
Control Group: Control
Alpha=0.050  Error Term=S(A)  DF=18  MSE=11541.3 Critical Value=2.3987

Treatment                                  Lower 95.0%   Difference   Upper 95.0%         Test
Group           Count            Mean     Simult.C.I.  With Control   Simult.C.I.         Result
T1                  7          660.43           -5.17        132.57        270.31
T2                  7          649.14          -16.46        121.29        259.03
```

The significance level (Prob Level) was only 0.061—not enough for statistical significance. Since the lower confidence limits are negative and the upper confidence limits are positive, Dunnett's two-sided test did not find a significant difference between either treatment and the control group.

The researcher had hoped to show that the treatment groups had higher response levels than the control group. He could see that the group means followed this pattern since the mean for *T1* was about 25% higher than the control mean and the mean for *T2* was about 23% higher than the control mean. He decided to calculate the power of the experiment.

Setup

The data entry for this problem is simple. The only entry that is not straight forward is finding an appropriate value for the standard deviation. Since the standard deviation is estimated by the square root of the mean square error, it is calculated as $\sqrt{11541.3} = 107.4304$.

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Comparisons (Analytic)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**7**
k (Number of Groups)	**3**
Group Sample Size Pattern	**Equal**
Minimum Detectable Difference	**133**
S (Standard Deviation of Subjects)	**107.4304**
Type of Multiple Comparison	**With Control - Dunnett**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
              Average                              Minimum    Standard
                 Size          Total            Detectable   Deviation
Power             (n)    k        N   Alpha   Beta  Difference        (S)       Diff / S
0.0002           7.00    3       21  0.0500 0.9998      133.00     107.43       1.2380
```

The power is only 0.0002. Hence, there was little chance of detecting a difference of 133 between a treatment and a control group.

It was of interest to the researcher to determine how large of a sample was needed if the power was to be 0.90. Setting Beta equal to 0.90 and ‘Find/Solve For’ to *n* resulted in the following report.

```
              Average                              Minimum    Standard
                 Size          Total            Detectable   Deviation
Power             (n)    k        N   Alpha   Beta  Difference        (S)       Diff / S
0.9042          33.00    3       99  0.0500 0.0958      133.00     107.43       1.2380
```

We see that instead of 7 per group, 33 per group were needed.

It was also of interest to the research to determine how large of a difference between the means could have been detected. Setting ‘Find’ to Min Detectable Difference resulted in the following report.

```
              Average                              Minimum    Standard
                 Size          Total            Detectable   Deviation
Power             (n)    k        N   Alpha   Beta  Difference        (S)       Diff / S
0.9000           7.00    3       21  0.0500 0.1000      348.81     107.43       3.2468
```

We see that a study of this size with these parameters could only detect a difference of 348.8. This explains why the results were not significant.

Example 3 – Using Unequal Sample Sizes

It is usually advisable to design experiments with equal sample sizes in each group. In some cases, however, it may be necessary to allocate subjects unequally across the groups. This may occur when the group variances are unequal, the costs per subject are different, or the dropout rates are different. This module can be used to study the power of unbalanced experiments.

In this example which will use Dunnett’s test, the minimum detectable difference is 2.0, the standard deviation is 1.0, alpha is 0.05, and *k* is 3. The sample sizes are 7, 7, and 14.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Comparisons (Analytic)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3a** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
n (Sample Size Multiplier)	**1**
k (Number of Groups)	**3**
Group Sample Size Pattern	**7 7 14**
Minimum Detectable Difference	**2**
S (Standard Deviation of Subjects)	**1**
Type of Multiple Comparison	**With Control - Dunnett**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
          Average                                  Minimum   Standard
             Size           Total                Detectable  Deviation
Power         (n)     k         N    Alpha   Beta  Difference       (S)        Diff / S
0.2726       9.33     3        28   0.0500  0.7274       2.00      1.00        2.0000
```

Alternatively, this problem could have been set up as follows (**Example3b** template):

Option	Value
Data Tab	
n (Sample Size Multiplier)	**7**
Group Sample Size Pattern	**1 1 2**

The advantage of this method is that you can try several values of *n* while keeping the same allocation ratios.

Example 4 – Validation using Hsu

Hsu (1996) page 241 presents an example of determining the sample size in an experiment with 8 groups. The minimum detectable difference is 10,000 psi. The standard deviation is 3,000 psi. Alpha is 0.05 and beta is 0.10. He finds a sample size of 10 per group for the Tukey-Kramer test, a sample size of 6 for Hsu’s test, and a sample size of 8 for Dunnett’s test.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Comparisons (Analytic)**. You may then follow along here by making the appropriate entries as listed below or load the completed templates **Example4**, **Example4b**, and **Example4c** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**n**
Power	**0.9**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
k (Number of Groups)	**8**
Group Sample Size Pattern	**Equal**
Minimum Detectable Difference	**10000**
S (Standard Deviation of Subjects)	**3000**
Type of Multiple Comparison	**All Pairs - Tukey Kramer**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results for Tukey-Kramer Test

	Average					Minimum	Standard	
	Size		Total			Detectable	Deviation	
Power	(n)	k	N	Alpha	Beta	Difference	(S)	Diff / S
0.9397	10.00	8	80	0.0500	0.0603	10000.00	3000.00	3.3333

PASS also found *n* = 10.

Numeric Results for Hsu’s Test

	Average					Minimum	Standard	
	Size		Total			Detectable	Deviation	
Power	(n)	k	N	Alpha	Beta	Difference	(S)	Diff / S
0.9087	6.00	8	48	0.0500	0.0913	10000.00	3000.00	3.3333

PASS also found *n* = 6.

Numeric Results for Dunnett’s Test

	Average					Minimum	Standard	
	Size		Total			Detectable	Deviation	
Power	(n)	k	N	Alpha	Beta	Difference	(S)	Diff / S
0.9434	8.00	8	64	0.0500	0.0566	10000.00	3000.00	3.3333

PASS found *n* = 8.

Example 5 – Validation using Pan and Kupper

Pan and Kupper (1999, page 1481) present examples of determining the sample size using alternative methods. It is interesting to compare the method of Hsu (1996) with theirs. Although the results are not exactly the same, they are very close.

In the example of Pan and Kupper, the minimum detectable difference is 0.50. The standard deviation is 0.50. Alpha is 0.05, and beta is 0.10. They find a sample size of 51 per group for Dunnett’s test and a sample size of 60 for the Tukey-Kramer test.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Comparisons (Analytic)**. You may then follow along here by making the appropriate entries as listed below or load the completed templates **Example5a** and **Example5b** from the Template tab on the procedure window.

tab.

Option	Value
Data Tab	
Find (Solve For)	**n**
Power	**0.9**
Alpha	**0.05**
n (Sample Size Multiplier)	*Ignored since this is the Find setting*
k (Number of Groups)	**4**
Group Sample Size Pattern	**Equal**
Minimum Detectable Difference	**0.50**
S (Standard Deviation of Subjects)	**0.50**
Type of Multiple Comparison	**With Control - Dunnett**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results for Dunnett’s Test

```
            Average                                          Minimum    Standard
               Size            Total                      Detectable   Deviation
 Power          (n)      k         N     Alpha      Beta  Difference         (S)        Diff / S
0.9146        53.00      4       212    0.0500    0.0854        0.50        0.50          1.0000
```

PASS found *n* = 53. This is very close to the 51 that Pan and Kupper found using a slightly different method.

Numeric Results for Tukey-Kramer Test

	Average					Minimum	Standard	
	Size		Total			Detectable	Deviation	
Power	(n)	k	N	Alpha	Beta	Difference	(S)	Diff / S
0.9057	62.00	4	248	0.0500	0.0943	0.50	0.50	1.0000

PASS also found *n* = 62. This is very close to the 60 that Pan and Kupper found using a slightly different method.

Chapter 580

Pair-Wise Multiple Comparisons (Simulation)

Introduction

This procedure uses simulation analyze the power and significance level of three pair-wise multiple-comparison procedures: Tukey-Kramer, Kruskal-Wallis, and Games-Howell. For each scenario, two simulations are run: one estimates the significance level and the other estimates the power.

The term *multiple comparisons* refers to a set of two or more statistical hypothesis tests. The term *pair-wise multiple comparisons* refers to the set of all pairs of means that can be generated among the means of *k* groups. For example, suppose the levels of a factor with five groups are labeled A, B, C, D, and E. The ten possible paired-comparisons that could be made among the five groups are A-B, A-C, A-D, A-E, B-C, B-D, B-E, C-D, C-E, and D-E.

As the number of groups increases, the number of comparisons (pairs) increases dramatically. For example, a 5 group design has 10 pairs, a 10 group design has 45 pairs, and a 20 group design has 190 pairs. When several comparisons are made among the group means, the determination of the significance level of each individual comparison is much more complex because of the problem of *multiplicity*. *Multiplicity* here refers to the fact that the chances of making at least one incorrect decision increases as the number of statistical tests increases. The method of *multiple comparisons* has been developed to account for this multiplicity.

Error Rates

When dealing with several simultaneous statistical tests, both individual-wise and experiment wise error rates should be considered.

1. **Comparison-wise error rate**. This is the probability of a type-I error (rejecting a true H0) for a particular test. In the case of the five-group design, there are ten possible comparison-wise error rates, one for each of the ten possible pairs. We will denote this error rate α_c.

2. **Experiment-wise (or family-wise) error rate**. This is the probability of making one or more type-I errors in the set (family) of comparisons. We will denote this error rate α_f.

The relationship between these two error rates when the tests are independent is given by

$$\alpha_f = 1 - (1 - \alpha_c)^C$$

where C is the total number of comparisons in the family. For example, if α_c is 0.05 and C is 10, α_f is 0.401. There is about a 40% chance that at least one of the ten pairs will be concluded to be different when in fact they are all the same. When the tests are correlated, as they are among a set of pair-wise comparisons, the above formula provides an upper bound to the family-wise error rate.

The techniques described below provide control for α_f rather than α_c.

Technical Details

The One-Way Analysis of Variance Design

The discussion that follows is based on the common one-way analysis of variance design which may be summarized as follows. Suppose the responses Y_{ij} in k groups each follow a normal distribution with means $\mu_1, \mu_2, \cdots, \mu_k$ and unknown variance σ^2. Let $n_1, n_2, \cdots, n_k$ denote the number of subjects in each group.

The analysis of these responses is based on the sample means

$$\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

and the pooled sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2}{\sum_{i=1}^{k} \left(n_i - 1\right)}$$

The F test is the usual method of analysis of the data from such a design, testing whether all of the means are equal. However, a significant F test does not indicate which of the groups are different, only that at least one is different. The analyst is left with the problem of determining which of the groups are different and by how much.

The Tukey-Kramer procedure, the Kruskal-Wallis procedure, and the Games-Howell procedure are the pair-wise multiple-comparison procedures that have been developed for this situation. The calculation of each of these tests is given next.

Tukey-Kramer

This test is referenced in Kirk (1982). It uses the critical values from the studentized-range distribution. For each pair of groups, the significance test between any two groups *i* and *j* is calculated by rejecting the null hypothesis of mean equality if

$$\frac{\left|\overline{Y}_i - \overline{Y}_j\right|}{\sqrt{\frac{\hat{\sigma}^2}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \geq q_{\alpha_f,k,v}$$

where

$$\begin{aligned} v &= \sum_{i=1}^{k} n_i - k \\ &= N - k \end{aligned}$$

Kruskal-Wallis

This test is attributed to Dunn (1964) and is referenced in Gibbons (1976). It is a nonparametric, or distribution-free, test for which the assumption of normality is not necessary. It tests whether pairs of medians are equal using a rank test. Sample sizes of at least five (but preferably larger) for each treatment are recommended for use of this test. The error rate is adjusted on a comparison-wise basis to give the experiment-wise error rate, α_f. Instead of using means, it uses average ranks, as the following formula indicates, with $\alpha = \alpha_f / \left(k(k-1)\right)$. For each pair of groups, *i* and *j*, the null hypothesis of equality is rejected if

$$\frac{|\overline{\mathrm{R}}_{\mathrm{i}} - \overline{\mathrm{R}}_{\mathrm{j}}|}{\sqrt{\frac{\mathrm{n(n+1)}}{12}\left(\frac{1}{\mathrm{n_i}} + \frac{1}{\mathrm{n_j}}\right)}} \geq \mathrm{z}_\alpha$$

Note that, when necessary, the usual adjustment for ties is made.

Games-Howell

This test is referenced in Kirk (1982) page 120. It was developed for the case when the individual group variances cannot be assumed to be equal. It also uses critical values from the studentized-range distribution. For each pair of groups, *i* and *j*, the null hypothesis of equality is rejected if

$$\frac{\left|\overline{Y}_i - \overline{Y}_j\right|}{\sqrt{\frac{1}{2}\left(\frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j}\right)}} \geq q_{\alpha_f,k,v'}$$

where

$$v' = \frac{\left(\frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j}\right)^2}{\frac{\hat{\sigma}_i^4}{n_i^2\left(n_i - 1\right)} + \frac{\hat{\sigma}_j^4}{n_j^2\left(n_j - 1\right)}}$$

If any of the following conditions hold, then $v' = n_i + n_2 - 2$:

1. $9/10 \le n_i / n_j \le 10/9$
2. $9/10 \le \left(\frac{\hat{\sigma}_i^2}{n_i}\right) / \left(\frac{\hat{\sigma}_j^2}{n_j}\right) \le 10/9$
3. $4/5 \le n_i / n_j \le 5/4$ and $1/2 \le \left(\frac{\hat{\sigma}_i^2}{n_i}\right) / \left(\frac{\hat{\sigma}_j^2}{n_j}\right) \le 2$
4. $2/3 \le n_i / n_j \le 3/2$ and $3/4 \le \left(\frac{\hat{\sigma}_i^2}{n_i}\right) / \left(\frac{\hat{\sigma}_j^2}{n_j}\right) \le 4/3$

Definition of Power for Multiple Comparisons

The notion of the power of a test is well-defined for individual tests. Power is the probability of rejecting a false null hypothesis. However, this definition does not extend easily when there are a number of simultaneous tests.

To understand the problem, consider an experiment with three groups labeled, A, B, and C. There are three paired comparisons in this experiment: A-B, A-C, and B-C. How do we define power for these three tests? One approach would be to calculate the power of each of the three tests, ignoring the other two. However, this ignores the interdependence among the three tests. Other definitions of the power of the set of tests might be the probability of detecting at least one of the differing pairs, exactly one of the differing pairs, at least two of the differing pairs, and so on. As the number of pairs increases, the number of possible definitions of power also increases. The two definitions that we emphasize in ***PASS*** were recommended by Ramsey (1978). They are *any-pair power* and *all-pairs power*. Other design characteristics, such as average-comparison power and false-discovery rate, are important to consider. However, our review of the statistical literature resulted in our focus on these two definitions of power.

Any-Pair Power

Any-pair power is the probability of detecting at least one of the pairs that are actually different.

All-Pairs Power

All-pairs power is the probability of detecting all of the pairs that are actually different.

Simulation Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify how each test is to be carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.
2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the power of each test.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the significance level of each test.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in ***PASS*** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draws the random numbers from this pool. This second method can cut the running time of the simulation by 70%!

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data, Reports, and Options tabs. To find out more about using the other tabs such as Axes/Legend, Plot Text, or Template, go to the Procedure Window chapter.

Data 1 Tab

The Data 1 tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for: power or sample size (n). If you choose to solve for n, you must choose the type of power you want to solve for: any-pair power or all-pairs power. The value of the option *Power* will then represent this type of power.

Any-pair power is the probability of detecting at least one pair from among those that are actually different. *All-pairs power* is the probability of detecting all pairs that are actually different.

Note that the search for n may take several minutes because a separate simulation must be run for each trial value of n. You may find it quicker and more informative to solve for the Power for a range of sample sizes.

Error Rates

Power or Beta

This option is only used when *Find (Solve For)* is set to *n (All-Pairs)* or *n (Any-Pair)*.

Power is defined differently with multiple comparisons. Although many definitions are possible, two are adopted here. *Any-pair power* is the probability of detecting at least one pair of the means that are different. *All-pairs power* is the probability of detecting all pairs of means that are truly different. As the number of groups is increased, these power probabilities will decrease because more tests are being conducted.

Since this is a probability, the range is between 0 and 1. Most researchers would like to have the power at least at 0.8. However, this may require extremely large sample sizes when the number of tests is large.

FWER (Alpha)

This option specifies one or more values of the *family-wise error rate* (FWER) which is the analog of alpha for multiple comparisons. FWER is the probability of falsely detecting (concluding that the means are different) at least one comparison for which the true means are the same. For independent tests, the relationship between the individual-comparison error rate (ICER) and FWER is given by the formulas

$$FWER = 1 - (1 - ICER)^{\wedge}C$$

or

$$ICER = 1 - (1 - FWER)^{\wedge}(1/C)$$

where '^' represents exponentiation (as in 4^2 = 16) and C represents the number of comparisons. For example, if C = 5 and FWER = 0.05, then ICER = 0.0102. Thus, the individual comparison tests must be conducted using a Type-1 error rate of 0.0102, which is much lower than the family-wise rate of 0.05.

The popular value for FWER remains at 0.05. However, if you have a large number of comparisons, you might decide that a larger value, such as 0.10, is appropriate.

Sample Size

n (Sample Size Multiplier)

This is the base, per group, sample size. One or more values separated by blanks or commas may be entered. A separate analysis is performed for each value listed here.

The group samples sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by *m1*, *m2*, *m3*, …, *mk* and this value is represented by *n*, the group sample sizes *N1*, *N2*, *N3*, ..., *Nk* are calculated as follows:

N1=[n(m1)]

N2=[n(m2)]

N3=[n(m3)]

etc.

where the operator, [*X*] means the next integer after *X*, e.g. [3.1]=4.

For example, suppose there are three groups and the Group Sample Size Pattern is set to *1,2,3*. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to *2,4,6,8,10*, five sets of group sample sizes will be generated and an analysis run for each. These sets are:

2	4	6
4	8	12
6	12	18
8	16	24
10	20	30

As a second example, suppose there are three groups and the Group Sample Size Pattern is *0.2,0.3,0.5*. When the fractional Pattern values sum to one, n can be interpreted as the total sample size of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

(0.2)12 = 2.4 which is rounded up to the next whole integer, 3.

(0.3)12 = 3.6 which is rounded up to the next whole integer, 4.

(0.5)12 = 6.

Note that in this case, 3+4+6 does not equal n (which is 12). This can happen because of rounding.

Group Sample Size Pattern

The purpose of the group sample size pattern is to allow several groups with the same sample size to be generated without having to type each individually.

A set of positive, numeric values (one for each row of distributions) is entered here. Each item specified in this list applies to the whole row of distributions. For example, suppose the entry is *1 2 1* and Grps 1 = 3, Grps 2 = 1, Grps 3 = 2. The sample size pattern used would be *1 1 1 2 1 1*.

The sample size of group *i* is found by multiplying the ith number from this list by the value of *n* and rounding up to the next whole number. The number of values must match the number of groups, *g*. When too few numbers are entered, 1’s are added. When too many numbers are entered, the extras are ignored.

- **Equal**

  If all sample sizes are to be equal, enter *Equal* here and the desired sample size in *n*. A set of *g* 1's will be used. This will result in *n1* = *n2* = … = *ng* = *n*. That is, all sample sizes are equal to *n*.

Test

MC Procedure

Specify which pair-wise multiple comparison procedure is to be reported from the simulations. The choices are

- **Tukey-Kramer**

  This is the most popular and the most often recommended.

- **Kruskal-Wallis**

  This is recommended when a nonparametric procedure is wanted.

- **Games-Howell**

  This is recommended when the variances of the groups are unequal.

Simulations

Simulations

This option specifies the number of iterations, *M*, used in the simulation. As the number of iterations is increased, the running time and accuracy are increased as well.

The precision of the simulated power estimates are calculated using the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the ‘Precision’ amount given in the table.

Simulation Size M	Precision when Power = 0.50	Precision when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

These options specify the distributions to be used in the two simulations, one set per row. The first option specifies the number of groups represented by the two distributions that follow. The second option specifies the distribution to be used in simulating the null hypothesis to determine the significance level (alpha). The third option specifies the distribution to be used in simulating the alternative hypothesis to determine the power.

Grps [1 – 3] (Grps 4 – 9 are found on the Data 2 tab)

This value specifies the number of groups specified by the H0 and H1 distribution statements to the right. Usually, you will enter ‘1’ to specify a single H0 and a single H1 distribution, or you will enter ‘0’ to indicate that the distributions specified on this line are to be ignored. This option lets you easily specify many identical distributions with a single phrase.

The total number of groups *g* is equal to the sum of the values for the three rows of distributions shown under the Data1 tab and the six rows of distributions shown under the Data2 tab.

Note that each item specified in the *Group Sample Size Pattern* option applies to the whole row of entries here. For example, suppose the *Group Sample Size Pattern* was *1 2 1* and Grps 1 = 3, Grps 2 = 1, and Grps 3 = 2. The sample size pattern would be *1 1 1 2 1 1*.

Note that since the first group is the control group, the value for Grps 1 is usually set to one.

Group Distribution(s)|H0

This entry specifies the distribution of one or more groups under the null hypothesis, H0. The magnitude of the differences of the means of these distributions, which is often summarized as the standard deviation of the means, represents the magnitude of the mean differences specified under H0. Usually, the means are assumed to be equal under H0, so their standard deviation should be zero except for rounding.

These distributions are used in the simulations that estimate the actual significance level. They also specify the value of the mean under the null hypothesis, H0. Usually, these distributions will be identical. The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean is entered first.

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Group Distribution(s)|H1

Specify the distribution of this group under the alternative hypothesis, H1. This distribution is used in the simulation that determines the power. A fundamental quantity in a power analysis is the amount of variation among the group means. In fact, classical power analysis formulas, this variation is summarized as the standard deviation of the means.

The important point to realize is that you must pay particular attention to the values you give to the means of these distributions because they are fundamental to the interpretation of the simulation.

For convenience in specifying a range of values, the parameters of the distribution can be specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean, M1, is entered first.

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)

Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Equivalence Margin

Specify the largest difference for which means from different groups will be considered equal. When specifying group distributions, it is possible to end up with scenarios where some means are slightly different from each other, even though they are intended to be equivalent. This often happens when specifying distributions of different forms (e.g. normal and gamma) for different groups, where the means are intended to be the same. The parameters used to specify different distributions do not always result in means that are EXACTLY equal. This value lets you control how different means can be and still be considered equal.

This value is not used to specify the hypothesized mean differences of interest. The hypothesized differences are specified using the means (or parameters used to calculate means) for the null and alternative distributions.

This value should be MUCH smaller than the hypothesized mean differences.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for *M0* in the distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for *M1* in the distribution specifications given above. Although it can be used wherever you want, *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B, C)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values for each letter using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Reports Tab

The Reports tab contains settings about the format of the output.

Select Output – Numeric Reports

Show Various Reports & Plots

These options let you specify whether you want to generate the standard reports and plots.

Show Inc's & 95% C.I.

Checking this option causes an additional line to be printed showing a 95% confidence interval for both the power and actual alpha and half the width of the confidence interval (the increment).

Select Output – Plots

Show Comparative Reports & Plots

These options let you specify whether you want to generate reports and plots that compare the test statistics that are available.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Random Numbers

Random Number Pool Size

This is the size of the pool of values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Example 1 – Power at Various Sample Sizes

An experiment is being designed to investigate the variety of response when an experiment is replicated under five different conditions. Previous studies have shown that the standard deviation within a group is 3.0. Researchers want to detect a shift in the mean of 3.0 or more. To accomplish this, they set the means of the first four groups to zero and the mean of the fifth group to 3.0. They want to investigate sample sizes of 5, 10, 15, and 20 subjects per group.

Although they will conduct an F-test on the data, their primary analysis will be a set of Tukey-Kramer multiple comparison tests. They set the FWER to 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Pair-Wise Multiple Comparisons (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Pair-Wise Comparisons (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**5 10 15 20**
Group Sample Size Pattern	**Equal**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Grps 1	**4**
Group 1 Distribution(s) | H0	**N(M0 S)**
Group 1 Distribution(s) | H1	**N(M0 S)**
Grps 2	**1**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M0 S)**
Minimum Difference	**0.5**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**3**
S	**3**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Simulation Summary Report

```
Summary of Simulations of 5 Groups
MC Procedure: Tukey-Kramer M.C. Test

                 Group     Total
         Any-    Smpl.     Smpl.      All-   S.D. of   S.D. of
Sim.    Pairs     Size      Size     Pairs     Means      Data    Actual    Target
No.     Power        n         N     Power    Sm|H1     SD|H1      FWER      FWER       M0        M1        S
1       0.251      5.0        25     0.014       1.2       3.0     0.043     0.050      0.0       3.0      3.0
       (0.019)   [0.232    0.269]   (0.005)   [0.008    0.019]   (0.009)   [0.034   0.051]

2       0.562     10.0        50     0.077       1.2       3.0     0.052     0.050      0.0       3.0      3.0
       (0.022)   [0.540    0.583]   (0.012)   [0.065    0.088]   (0.010)   [0.042   0.062]

3       0.770     15.0        75     0.200       1.2       3.0     0.057     0.050      0.0       3.0      3.0
       (0.018)   [0.751    0.788]   (0.018)   [0.182    0.217]   (0.010)   [0.046   0.067]

4       0.898     20.0       100     0.356       1.2       3.0     0.060     0.050      0.0       3.0      3.0
       (0.013)   [0.884    0.911]   (0.021)   [0.335    0.377]   (0.010)   [0.049   0.070]


Pool Size: 10000. Simulations: 2000. Run Time: 26.45 seconds.

Summary of Simulations Report Definitions
H0: the null hypothesis that each pair of group means are equal.
H1: the alternative hypothesis that at least one pair of group means are not equal.
All-Pairs Power: the estimated probability of detecting all unequal pairs.
Any-Pairs Power: the estimated probability of detecting at least one unequal pair.
n: the average of the group sample sizes.
N: the combined sample size of all groups.
Family-Wise Error Rate (FWER): the probability of detecting at least one equal pair assuming H0.
Target FWER: the user-specified FWER.
Actual FWER: the FWER estimated by the alpha simulation.
Sm|H1: the standard deviation of the group means under H1.
SD|H1: the pooled, within-group standard deviation under H1.
Second Row: provides the precision and a confidence interval based on the size of the simulation for
Any-Pairs Power, All-Pairs Power, and FWER. The format is (Precision) [95% LCL and UCL Alpha].

Summary Statements
A one-way design with 5 groups has an average group sample size of 5.0 for a total sample size
of 25. This design achieved an any-pair power of 0.2505 and an all-pair power of 0.0135 using
the Tukey-Kramer M.C. Test with a target family-wise error rate of 0.050 and an actual target
family-wise error rate 0.043. The average within group standard deviation assuming the
alternative distribution is 3.0. These results are based on 2000 Monte Carlo samples from the
null distributions: N(M0 S); N(M0 S); N(M0 S); N(M0 S); and N(M0 S) and the alternative
distributions: N(M0 S); N(M0 S); N(M0 S); N(M0 S); and N(M1 S). Other parameters used in the
simulation were: M0 = 0.0, M1 = 3.0, and S = 3.0.
```

This report shows the estimated any-pairs power, all-pairs power, and FWER for each scenario. The second row shows three 95% confidence intervals in brackets: the first for the any-pairs power, the second for the all-pairs power, and the third for the FWER. Half the width of each confidence interval is given in parentheses as a fundamental measure of the precision of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Any-Pairs Power

This is the probability of detecting any of the significant pairs. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

All-Pairs Power

This is the probability of detecting all of the significant pairs. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

Group Sample Size n

This is the average of the individual group sample sizes.

Total Sample Size N

This is the total sample size of the study.

S.D. of Means Sm|H1

This is the standard deviation of the hypothesized means of the alternative distributions. Under the null hypothesis, this value is zero. This value represents the magnitude of the difference among the means that is being tested. It is roughly equal to the average difference between the group means and the overall mean.

Note that the effect size is the ratio of Sm|H1 and SD|H1.

S.D. of Data SD|H1

This is the within-group standard deviation calculated from samples from the alternative distributions.

Actual FWER

This is the value of FWER (family-wise error rate) estimated by the simulation using the H0 distributions. It should be compared with the Target FWER to determine if the test procedure is accurate.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the Actual FWER.

Target FWER

This is the target value of FWER that was set by the user.

M0

This is the value entered for M0, the group means under H0.

M1

This is the value entered for M1, the group means under H1.

S

This is the value entered for S, the standard deviation.

Error-Rate Summary for H0 Simulation

Error Rate Summary from H0 (Alpha) Simulation of 5 Groups
MC Procedure: Tukey-Kramer M.C. Test

Sim. No.	No. of Equal Pairs	Mean No. of Type-1 Errors	Prop. Type-1 Errors	Prop. (No. of Type-1 Errors > 0) FWER	Target FWER	Mean Pairs Alpha	Min Pairs Alpha	Max Pairs Alpha
1	10	0.057	0.006	0.043	0.050	0.006	0.005	0.007
2	10	0.072	0.007	0.052	0.050	0.007	0.005	0.011
3	10	0.075	0.008	0.057	0.050	0.008	0.006	0.011
4	10	0.075	0.008	0.060	0.050	0.008	0.004	0.010

This report shows the results of the H0 simulation. This simulation uses the H0 settings for each group. Its main purpose is to provide an estimate of the FWER.

No. of Equal Pairs

Since under H0 all means are equal, this is the number of unique pairs of the groups. Thus, this is the number of pair-wise multiple comparisons.

Mean No. of Type-1 Errors

This is the average number of type-1 errors (false detections) per set (family).

Prop. Type-1 Errors

This is the proportion of type-1 errors (false detections) among all tests that were conducted.

Prop. (No. of Type-1 Errors>0) FWER

This is the proportion of the H0 simulations in which at least one type-1 error occurred. This is called the family-wise error rate.

Target FWER

This is the target value of FWER that was set by the user.

Mean Pairs Alpha

Alpha is the probability of rejecting H0 when H0 is true. It is a characteristic of an individual test. This is the average alpha value over all of the tests in the family.

Min Pairs Alpha

This is the minimum of all of the individual comparison alphas.

Max Pairs Alpha

This is the maximum of all of the individual comparison alphas.

Error-Rate Summary for H1 Simulation

Error Rate Summary from H1 (Power) Simulation of 5 Groups
MC Procedure: Tukey-Kramer M.C. Test

Sim. No.	No. of Equal/ Uneq. Pairs	Mean No. of False Pos.	Mean No. of False Neg.	Prop. Errors	Prop. Equal that were Detect.	Prop. Uneq. that were Undet.	(FDR) Prop. Detect. that were Equal	Prop. Undet. that were Uneq.	All Uneq. Pairs Power	Any Uneq. Pairs Power	Mean Pairs Power	Min Pairs Power	Max Pairs Power
1	6/4	0.04	3.58	0.362	0.007	0.896	0.091	0.375	0.014	0.251	0.046	0.004	0.109
2	6/4	0.04	2.85	0.289	0.007	0.711	0.035	0.323	0.077	0.562	0.120	0.006	0.302
3	6/4	0.03	2.11	0.214	0.006	0.527	0.017	0.261	0.200	0.770	0.192	0.002	0.478
4	6/4	0.03	1.41	0.144	0.005	0.352	0.012	0.191	0.356	0.898	0.262	0.003	0.658

This report shows the results of the H1 simulation. This simulation uses the H1 settings for each group. Its main purpose is to provide an estimate of the power.

No. of Equal Pairs/Unequal Pairs

The first value is the number of pairs for which the means were equal under H1. The second value is the number of pairs for which the means were different under H1.

Mean No. False Positives

This is the average number of equal pairs that were declared as being unequal by the testing procedure. A *false positive* is a type-1 (alpha) error.

Mean No. False Negatives

This is the average number of unequal pairs that were not declared as being unequal by the testing procedure. A *false negative* is a type-2 (beta) error.

Prop. Errors

This is the proportion of type-1 and type-2 errors.

Prop. Equal that were Detect.

This is the proportion of the equal pairs in the H1 simulations that were declared as unequal.

Prop. Uneq. that were Undet.

This is the proportion of the unequal pairs in the H1 simulations that were not declared as being unequal.

Prop. Detect. that were Equal (FDR)

This is the proportion of all detected pairs in the H1 simulations that were actually equal. This is often called the *false discovery rate*.

Prop. Undet. that were Uneq.

This is the proportion of undetected pairs in the H1 simulations that were actually unequal.

All Uneq. Pairs Power

This is the probability of detecting all of the pairs that were different in the H1 simulation.

Any Uneq. Pairs Power

This is the probability of detecting any of the pairs that were different in the H1 simulation.

Mean, Min, and Max Pairs Power

These items give the average, the minimum, and the maximum of the individual comparison powers from the H1 simulation.

Detail Model Report

```
Detailed Model Report for Simulation No. 1
Target FWER = 0.050, M0 = 0.0, M1 = 3.0, S = 3.0
MC Procedure: Tukey-Kramer M.C. Test

Hypo.          Group                Group      Ave.    Simulation
Type   Groups  Labels      n/N       Mean      S.D.    Model
H0     1-4     A1-A4      5/25        0.0       2.9    N(M0 S)
H0     5       B1         5/25        0.0       3.0    N(M0 S)
H0     All                         Sm=0.0       3.0

H1     1-4     A1-A4      5/25        0.0       3.0    N(M0 S)
H1     5       B1         5/25        3.0       3.0    N(M1 S)
H1     All                         Sm=1.2       3.0

Detailed Model Report for Simulation No. 2
Hypo.          Group                Group      Ave.    Simulation
Type   Groups  Labels      n/N       Mean      S.D.    Model
H0     1-4     A1-A4     10/50        0.0       3.0    N(M0 S)
H0     5       B1        10/50        0.0       3.0    N(M0 S)
H0     All                         Sm=0.0       3.0

H1     1-4     A1-A4     10/50        0.0       3.0    N(M0 S)
H1     5       B1        10/50        3.0       3.0    N(M1 S)
H1     All                         Sm=1.2       3.0

Detailed Model Report for Simulation No. 3
Hypo.          Group                Group      Ave.    Simulation
Type   Groups  Labels      n/N       Mean      S.D.    Model
H0     1-4     A1-A4     15/75        0.0       3.0    N(M0 S)
H0     5       B1        15/75        0.0       3.0    N(M0 S)
H0     All                         Sm=0.0       3.0

H1     1-4     A1-A4     15/75        0.0       3.0    N(M0 S)
H1     5       B1        15/75        3.0       3.0    N(M1 S)
H1     All                         Sm=1.2       3.0

Detailed Model Report for Simulation No. 4
Hypo.          Group                Group      Ave.    Simulation
Type   Groups  Labels      n/N       Mean      S.D.    Model
H0     1-4     A1-A4    20/100        0.0       3.0    N(M0 S)
H0     5       B1       20/100        0.0       3.0    N(M0 S)
H0     All                         Sm=0.0       3.0

H1     1-4     A1-A4    20/100        0.0       3.0    N(M0 S)
H1     5       B1       20/100        3.0       3.0    N(M1 S)
H1     All                         Sm=1.2       3.0
```

This report shows details of each row of the previous reports.

Hypo. Type

This indicates which simulation is being reported on each row. H0 represents the null simulation and H1 represents the alternative simulation.

Groups

Each group in the simulation is assigned a number. This item shows the arbitrary group number that was assigned.

Group Labels

These are the labels that were used in the individual alpha-level reports.

n/N

n is the average sample size of the groups. N is the total sample size across all groups.

Group Mean

These are the means of the individual groups as specified for the H0 and H1 simulations.

Ave. S.D.

This is the average standard deviation of all groups reported on each line. Note that it is calculated from the simulated data.

Simulation Model

This is the distribution that was used to simulate data for the groups reported on each line.

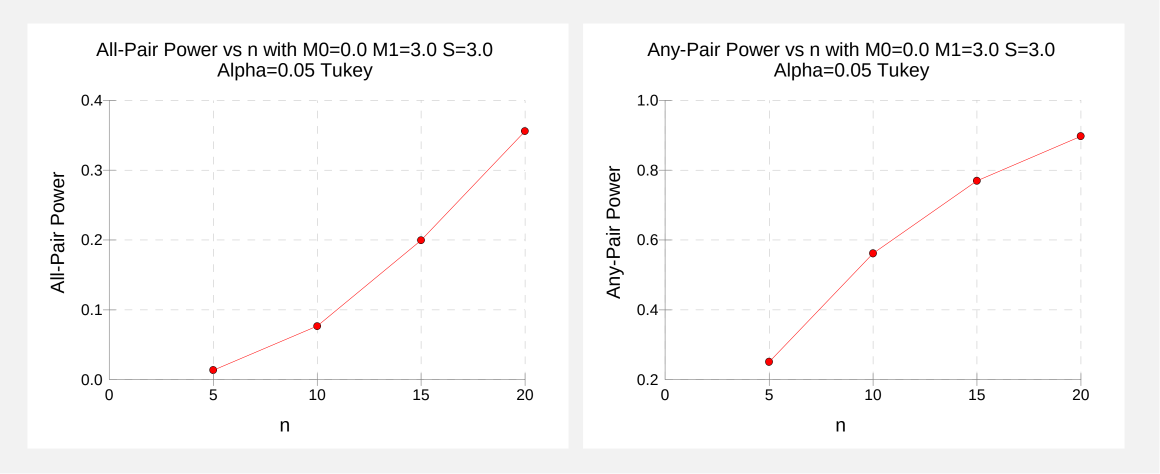
Probability of Rejecting Equality

```
Probability of Rejecting the Equality of Each Pair. Simulation No. 1

Group    Means      A1       A2       A3       A4       B1
A1         0.0             0.006    0.006    0.011   0.109*
A2         0.0    0.006             0.006    0.004   0.106*
A3         0.0    0.005    0.006             0.008   0.097*
A4         0.0    0.005    0.007    0.007            0.107*
B1         3.0    0.005    0.005    0.007    0.006

Probability of Rejecting the Equality of Each Pair. Simulation No. 2

Group    Means      A1       A2       A3       A4       B1
A1         0.0             0.007    0.007    0.008   0.294*
A2         0.0    0.006             0.007    0.006   0.280*
A3         0.0    0.011    0.007             0.006   0.302*
A4         0.0    0.005    0.007    0.005            0.281*
B1         3.0    0.008    0.009    0.008    0.007

Probability of Rejecting the Equality of Each Pair. Simulation No. 3

Group    Means      A1       A2       A3       A4       B1
A1         0.0             0.005    0.002    0.007   0.478*
A2         0.0    0.006             0.006    0.007   0.477*
A3         0.0    0.007    0.008             0.005   0.473*
A4         0.0    0.009    0.008    0.007            0.465*
B1         3.0    0.006    0.011    0.007    0.008

Probability of Rejecting the Equality of Each Pair. Simulation No. 4

Group    Means      A1       A2       A3       A4       B1
A1         0.0             0.003    0.005    0.004   0.645*
A2         0.0    0.008             0.008    0.007   0.650*
A3         0.0    0.006    0.008             0.004   0.640*
A4         0.0    0.010    0.007    0.009            0.658*
B1         3.0    0.010    0.007    0.004    0.007

Individual pairwise powers from the H1 (Power) simulation are shown in the upper-right section.
Individual pairwise significance levels from the H0 (Alpha) simulation are shown in the lower-left section.
* Starred values are the powers of pairs that are unequal under H1.
```

This report shows the individual probabilities of rejecting each pair. When a pair was actually different, the value is the power of that test. These power values are starred.

The results shown on the upper-right section of each simulation report are from the H1 simulation. The results shown on the lower-left section of the report are from the H0 simulation.

Plots Section



All-Pair Power vs n with M0=0.0 M1=3.0 S=3.0 Alpha=0.05 Tukey

Any-Pair Power vs n with M0=0.0 M1=3.0 S=3.0 Alpha=0.05 Tukey

These plots give a visual presentation of the all-pairs power values and the any-pair power values.

Example 2 – Comparative Results

Continuing with Example 1, the researchers want to study the characteristics of alternative multiple comparison procedures.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Pair-Wise Multiple Comparisons (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Pair-Wise Comparisons (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**5 10 15 20**
Group Sample Size Pattern	**Equal**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Grps 1	**4**
Group 1 Distribution(s) | H0	**N(M0 S)**
Group 1 Distribution(s) | H1	**N(M0 S)**

Data Tab (continued)

Grps 2..**1**
Group 2 Distribution(s) | H0 ....**N(M0 S)**
Group 2 Distribution(s) | H1 ....**N(M1 S)**
Minimum Difference **0.5**
M0 (Mean|H0) **0**
M1 (Mean|H1) **3**
S **3**

Reports Tab

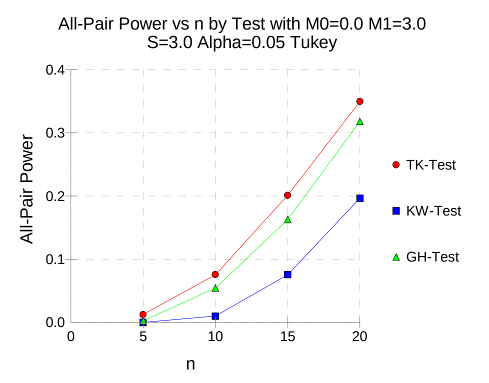
Comparative Reports ....**Checked**
Comparative Any-Pair Power Plot ....**Checked**
Comparative All-Pair Power Plot.............**Checked**

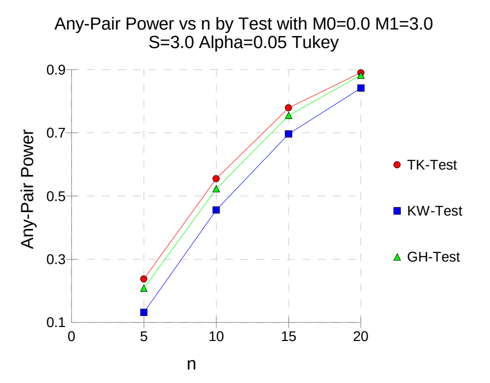
Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Power Comparison for Testing the 5 Group Means

                                  Tukey  Kruskal  Games    Tukey  Kruskal    Games
            Total                Kramer   Wallis Howell   Kramer   Wallis   Howell
Sim.       Sample    Target    All-Pair All-Pair All-Pair Any-Pair Any-Pair Any-Pair
No.          Size     Alpha      Power    Power    Power    Power    Power    Power
1              25     0.050      0.013    0.000    0.003    0.238    0.132    0.209
2              50     0.050      0.076    0.010    0.055    0.555    0.456    0.523
3              75     0.050      0.201    0.076    0.163    0.779    0.697    0.755
4             100     0.050      0.350    0.197    0.318    0.890    0.842    0.883
Pool Size: 10000. Simulations: 2000. Run Time: 33.82 seconds.

Family-Wise FWER Comparison for Testing the 5 Group Means

            Total                 Tukey  Kruskal   Games
Sim.       Sample    Target      Kramer   Wallis  Howell
No.          Size      FWER        FWER     FWER    FWER
1              25     0.050       0.039    0.023   0.064
2              50     0.050       0.050    0.039   0.059
3              75     0.050       0.051    0.033   0.055
4             100     0.050       0.051    0.040   0.058
```





These reports show the power and FWER of each of the three multiple comparison procedures. In these simulations of groups from the normal distributions with equal variances, we see that the Tukey-Kramer procedure is the champion.

Example 3 – Validation using Ramsey

Ramsey (1978) presents the results of a simulation study that compared the all-pair power of several different multiple comparison procedures. On page 483 of this article, he presents the results of a simulation in which there were four groups: two with means of -0.7 and two with means of 0.7. The standard deviation was 1.0 and the FWER was 0.05. Tukey's multiple comparison procedure was used in the simulation. The sample size was 16 per group. Using a simulation of 1000 iterations, the all-pairs power was calculated as 0.723. Note that a confidence interval for this estimated all-pairs power is (0.703 to 0.759).

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Pair-Wise Multiple Comparisons (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Pair-Wise Comparisons (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**16**
Group Sample Size Pattern	**Equal**
MC Procedure	**Tukey-Kramer**
Simulations	**2000**
Grps 1	**2**
Group 1 Distribution(s) | H0	**N(M0 S)**
Group 1 Distribution(s) | H1	**N(M0 S)**
Grps 2	**2**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M1 S)**
Minimum Difference	**0.1**
M0 (Mean|H0)	**-0.7**
M1 (Mean|H1)	**0.7**
S	**1**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Summary of Simulations of 5 Groups
MC Procedure: Tukey-Kramer M.C. Test

                Group     Total
        Any-    Smpl.     Smpl.      All-    S.D. of   S.D. of
Sim.   Pairs     Size      Size     Pairs     Means      Data    Actual    Target
No.    Power        n         N     Power     Sm|H1     SD|H1      FWER      FWER      M0       M1       S
1      0.996     16.0        64     0.729       0.7       1.0     0.052     0.050    -0.7      0.7     1.0
     (0.004)   [0.992    1.000]   (0.028)    [0.701    0.757]   (0.014)    [0.038  0.066]

Pool Size: 10000. Simulations: 1000. Run Time: 4.28 seconds.
```

Note that the value found by ***PASS*** of 0.729 is very close to the value of 0.723 found by Ramsey (1978). More importantly, the value found by ***PASS*** is inside the confidence limits of Ramsey's study.

We ran the simulation five more times and obtained 0.727, 0.723, 0.714, 0.740, and 0.732. We also ran the simulation with 10,000 iterations and obtained a power of 0.736 with a confidence interval of (0.727 to 0.745).

Example 4 – Selecting a Multiple Comparison Procedure when the Data Contain Outliers

This example will investigate the impact of outliers on the power and precision of the various multiple comparison procedures when there are five groups.

A mixture of two normal distributions will be used to randomly generate outliers. The mixture will draw 95% of the data from a normal distribution with mean zero and variance one. The other 5% of the data will come from a normal distribution with mean zero and variance that ranges from one to ten. In the alternative distributions, two will have means of zero and the other three will have means of one.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Pair-Wise Multiple Comparisons (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Pair-Wise Comparisons (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**10 20**

Data Tab (continued)

Group Sample Size Pattern	**Equal**
Multiple Comparison Procedure	**Tukey Kramer**
Simulations	**2000**
Grps 1	**2**
Group 1 Distribution(s) | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 1 Distribution(s) | H1	**N(M0 S)[95];N(M0 A)[5]**
Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Distribution(s) | H1	**N(M1 S)[95];N(M1 A)[5]**
Minimum Difference	**0.5**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**1**
S	**1**
A	**1 5 10**

Reports Tab

Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

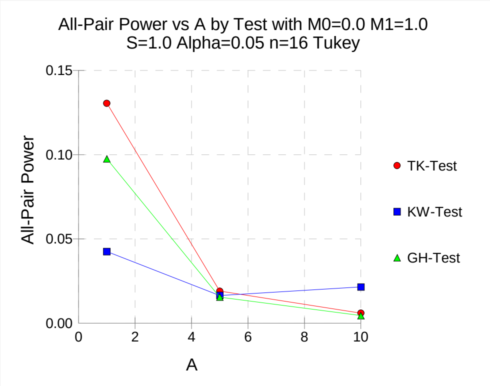
Power Comparison for Testing the 5 Group Means

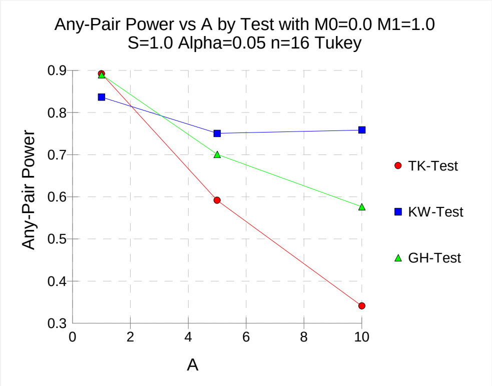
			Tukey	Kruskal	Games	Tukey	Kruskal	Games
	Total		Kramer	Wallis	Howell	Kramer	Wallis	Howell
Sim.	Sample	Target	All-Pair	All-Pair	All-Pair	Any-Pair	Any-Pair	Any-Pair
No.	Size	Alpha	Power	Power	Power	Power	Power	Power
1	80	0.050	0.131	0.043	0.098	0.892	0.837	0.890
2	80	0.050	0.019	0.017	0.016	0.592	0.751	0.701
3	80	0.050	0.006	0.022	0.005	0.341	0.759	0.577

Pool Size: 10000. Simulations: 2000. Run Time: 33.82 seconds.

Family-Wise Error-Rate Comparison for Testing the 5 Group Means

	Total		Tukey	Kruskal	Games
Sim.	Sample	Target	Kramer	Wallis	Howell
No.	Size	FWER	FWER	FWER	FWER
1	80	0.050	0.046	0.036	0.049
2	80	0.050	0.044	0.044	0.040
3	80	0.050	0.033	0.038	0.025





These reports show the power and FWER of each of the three multiple comparison procedures. We note that when the variances are equal (A = 1), the Tukey-Kramer procedure performs only slightly better than the others. However, as the number of outliers is increased, the Kruskal-Wallis procedure emerges as the better choice. Also note that in the case with many outliers (Simulation 3), the FWER of all procedures is below the target value.

Chapter 585

Multiple Comparisons of Treatments vs. a Control (Simulation)

Introduction

This procedure uses simulation to analyze the power and significance level of two multiple-comparison procedures that perform two-sided hypothesis tests of each treatment group mean versus the control group mean using simulation. These are Dunnett's test and the Kruskal-Wallis test. For each scenario, two simulations are run: one estimates the significance level and the other estimates the power.

The term *multiple comparisons of treatments versus a control* refers to the set of comparisons of each treatment group to a control group. If there are *k* groups of which *k*-1 are treatment groups, there will be *k*-1 tests.

When several comparisons are made among the group means, the interpretation for each comparison becomes more complex because of the problem of *multiplicity*. *Multiplicity* here refers to the fact that the chances of making at least one incorrect decision increases as the number of statistical tests increases. The method of *multiple comparisons* has been developed to account for this multiplicity.

Error Rates

When dealing with several simultaneous statistical tests, both individual-wise and experiment wise error rates should be considered.

1. **Comparison-wise error rate**. This is the probability of a type-I error (rejecting a true H0) for a particular test. In the case of the five-group design, there are ten possible comparison-wise error rates, one for each of the ten possible pairs. We will denote this error rate α_c.
2. **Experiment-wise (or family-wise) error rate**. This is the probability of making one or more type-I errors in the set (family) of comparisons. We will denote this error rate α_f.

The relationship between these two error rates when the tests are independent is given by

$$\alpha_f = 1 - (1 - \alpha_c)^C$$

where *C* is the total number of comparisons in the family. For example, if α_c is 0.05 and *C* is 10, α_f is 0.401. There is about a 40% chance that at least one of the ten pairs will be concluded to be different when in fact they are all the same. When the tests are correlated, as they are among a set of pair-wise comparisons, the above formula provides an upper bound to the family-wise error rate.

The techniques described below provide control for α_f rather than α_c.

Technical Details

The One-Way Analysis of Variance Design

The discussion that follows is based on the common one-way analysis of variance design which may be summarized as follows. Suppose the responses Y_{ij} in *k* groups each follow a normal distribution with means $\mu_1, \mu_2, \cdots, \mu_k$ and unknown variance σ^2. Let $n_1, n_2, \cdots, n_k$ denote the number of subjects in each group. The control group is assumed to be group one.

The analysis of these responses is based on the sample means

$$\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

and the pooled sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2}{\sum_{i=1}^{k} \left(n_i - 1\right)}$$

The *F* test is the usual method of analysis of the data from such a design, testing whether all of the means are equal. However, a significant *F* test does not indicate which of the groups are different, only that at least one is different. The analyst is left with the problem of determining which of the groups are different and by how much.

The Dunnett procedure and a special version of the Kruskal-Wallis procedure have been developed for this situation. The calculation of each of these tests is given next.

Dunnett's Test

Dunnett (1955) developed a test procedure for simultaneously comparing each treatment with a control group. It uses the critical values of a special *t* distribution given in Dunnett (1955). For each treatment and control pair, the significance test is calculated by rejecting the null hypothesis of mean equality if

$$\frac{\left|\overline{Y}_i - \overline{Y}_1\right|}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_i}+\frac{1}{n_1}\right)}} \geq t_{1-\alpha/2,k,N-k}^{Dunnett} \qquad i = 2,\ldots,k$$

Kruskal-Wallis Test

This test is attributed to Dunn (1964) and is referenced in Gibbons (1976). It is a nonparametric, or distribution-free, test which means that normality is not assumed. It tests whether the medians of each treatment-control pair are equal using a rank test. Sample sizes of at least five (but preferably larger) for each treatment are recommended for the use of this test. The error rate is adjusted on a comparison-wise basis to give the experiment-wise error rate, α_f. Instead of using means, it uses average ranks as the following formula indicates, with $\alpha = \alpha_f / \left(2\left(k-1\right)\right)$. For each treatment and control pair, the significance test is calculated by rejecting the null hypothesis of median equality if

$$\frac{\left|\overline{R}_i - \overline{R}_1\right|}{\sqrt{\frac{N\left(N+1\right)}{12}\left(\frac{1}{n_i}+\frac{1}{n_1}\right)}} \geq z_\alpha \qquad i = 2,\ldots,k$$

Note that, when necessary, the usual adjustment for ties is made.

Definition of Power for Multiple Comparisons

The notion of the power of a test is well-defined for individual tests. Power is the probability of rejecting a false null hypothesis. However, this definition does not extend easily when there are a number of simultaneous tests.

To understand the problem, consider an experiment with four groups labeled, C, A, B, and D. Suppose C is the control group. There are three paired comparisons in this experiment: A-C, B-C, and D-C. How do we define power for these three tests? One approach would be to calculate the power of each of the three tests, ignoring the other two. However, this ignores the interdependence among the three tests. Other definitions of the power of the set of tests might be the probability of detecting at least one of the differing pairs, exactly one of the differing pairs, at least two of the differing pairs, and so on. As the number of pairs increases, the number of possible definitions of power also increases. The two definitions that we emphasize in ***PASS*** were recommended by Ramsey (1978). They are *any-pair power* and *all-pairs power*. Other design characteristics, such as average-comparison power and false-discovery rate, are important to consider. However, our review of the statistical literature resulted in our focus on these two definitions of power.

Any-Pair Power

Any-pair power is the probability of detecting at least one of the pairs that are actually different.

All-Pairs Power

All-pairs power is the probability of detecting all of the pairs that are actually different.

Simulation Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify how each test is to be carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.
2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the power of each test.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the significance level of each test.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in ***PASS*** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draws the random numbers from this pool. This second method can cut the running time of the simulation by 70%!

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data, Reports, and Options tabs. To find out more about using the other tabs such as Axes/Legend, Plot Text, or Template, go to the Procedure Window chapter.

Data 1 Tab

The Data 1 tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for: power or sample size (n). If you choose to solve for n, you must choose the type of power you want to solve for: any-pair power or all-pairs power. The value of the option *Power* will then represent this type of power.

Any-pair power is the probability of detecting at least one pair from among those that are actually different. *All-pairs power* is the probability of detecting all pairs that are actually different.

Note that the search for n may take several minutes because a separate simulation must be run for each trial value of n. You may find it quicker and more informative to solve for the Power for a range of sample sizes.

Error Rates

Power or Beta

This option is only used when *Find (Solve For)* is set to *n (All-Pairs)* or *n (Any-Pair)*.

Power is defined differently with multiple comparisons. Although many definitions are possible, two are adopted here. *Any-pair power* is the probability of detecting at least one pair of the means that are different. *All-pairs power* is the probability of detecting all pairs of means that are truly different. As the number of groups is increased, these power probabilities will decrease because more tests are being conducted.

Since this is a probability, the range is between 0 and 1. Most researchers would like to have the power at least at 0.8. However, this may require extremely large sample sizes when the number of tests is large.

FWER (Alpha)

This option specifies one or more values of the *family-wise error rate* (FWER) which is the analog of alpha for multiple comparisons. FWER is the probability of falsely detecting (concluding that the means are different) at least one comparison for which the true means are the same. For independent tests, the relationship between the individual-comparison error rate (ICER) and FWER is given by the formulas

$$FWER = 1 - (1 - ICER)\hat{}C$$

or

$$ICER = 1 - (1 - FWER)\text{^}(1/C)$$

where '^' represents exponentiation (as in 4^2 = 16) and C represents the number of comparisons. For example, if C = 5 and FWER = 0.05, then ICER = 0.0102. Thus, the individual comparison tests must be conducted using a Type-1 error rate of 0.0102, which is much lower than the family-wise rate of 0.05.

The popular value for FWER remains at 0.05. However, if you have a large number of comparisons, you might decide that a larger value, such as 0.10, is appropriate.

Sample Size

n (Sample Size Multiplier)

This is the base, per group, sample size. One or more values separated by blanks or commas may be entered. A separate analysis is performed for each value listed here.

The group samples sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by *m1*, *m2*, *m3*, …, *mk* and this value is represented by *n*, the group sample sizes *N1*, *N2*, *N3*, ..., *Nk* are calculated as follows:

N1=[n(m1)]

N2=[n(m2)]

N3=[n(m3)]

etc.

where the operator, [*X*] means the next integer after *X*, e.g. [3.1]=4.

For example, suppose there are three groups and the Group Sample Size Pattern is set to *1,2,3*. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to *2,4,6,8,10*, five sets of group sample sizes will be generated and an analysis run for each. These sets are:

2	4	6
4	8	12
6	12	18
8	16	24
10	20	30

As a second example, suppose there are three groups and the Group Sample Size Pattern is *0.2,0.3,0.5*. When the fractional Pattern values sum to one, n can be interpreted as the total sample size of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

(0.2)12 = 2.4 which is rounded up to the next whole integer, 3.

(0.3)12 = 3.6 which is rounded up to the next whole integer, 4.

(0.5)12 = 6.

Note that in this case, 3+4+6 does not equal n (which is 12). This can happen because of rounding.

Group Sample Size Pattern

The purpose of the group sample size pattern is to allow several groups with the same sample size to be generated without having to type each individually.

A set of positive, numeric values (one for each row of distributions) is entered here. Each item specified in this list applies to the whole row of distributions. For example, suppose the entry is *1 2 1* and Grps 1 = 3, Grps 2 = 1, Grps 3 = 2. The sample size pattern used would be *1 1 1 2 1 1*.

The sample size of group *i* is found by multiplying the ith number from this list by the value of *n* and rounding up to the next whole number. The number of values must match the number of groups, *g*. When too few numbers are entered, 1's are added. When too many numbers are entered, the extras are ignored.

- **Equal**

  If all sample sizes are to be equal, enter *Equal* here and the desired sample size in *n*. A set of *g* 1's will be used. This will result in *n1* = *n2* = … = *ng* = *n*. That is, all sample sizes are equal to *n*.

Test

MC Procedure

Specify which pair-wise multiple comparison procedure is to be reported from the simulations. The choices are

- **Dunnett Test**

  This is the most popular and the most often recommended.

- **Kruskal-Wallis**

  This is recommended when a nonparametric procedure is wanted.

Simulations

Simulations

This option specifies the number of iterations, *M*, used in the simulation. As the number of iterations is increased, the running time and accuracy are increased as well.

The precision of the simulated power estimates are calculated using the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the 'Precision' amount given in the table.

Simulation Size M	Precision when Power = 0.50	Precision when Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

These options specify the distributions to be used in the two simulations, one set per row. The first option specifies the number of groups represented by the two distributions that follow. The second option specifies the distribution to be used in simulating the null hypothesis to determine the significance level (alpha). The third option specifies the distribution to be used in simulating the alternative hypothesis to determine the power.

Note that group number one is the control group.

Grps [1 – 3] (Grps 4 – 9 are found on the Data 2 tab)

This value specifies the number of groups specified by the H0 and H1 distribution statements to the right. Usually, you will enter ‘1’ to specify a single H0 and a single H1 distribution, or you will enter ‘0’ to indicate that the distributions specified on this line are to be ignored. This option lets you easily specify many identical distributions with a single phrase.

The total number of groups *g* is equal to the sum of the values for the three rows of distributions shown under the Data1 tab and the six rows of distributions shown under the Data2 tab.

Note that each item specified in the *Group Sample Size Pattern* option applies to the whole row of entries here. For example, suppose the *Group Sample Size Pattern* was *1 2 1* and Grps 1 = 3, Grps 2 = 1, and Grps 3 = 2. The sample size pattern would be *1 1 1 2 1 1*.

Note that since the first group is the control group, the value for Grps 1 is usually set to one.

Group Distribution(s)|H0

This entry specifies the distribution of one or more groups under the null hypothesis, H0. The magnitude of the differences of the means of these distributions, which is often summarized as the standard deviation of the means, represents the magnitude of the mean differences specified under H0. Usually, the means are assumed to be equal under H0, so their standard deviation should be zero except for rounding.

These distributions are used in the simulations that estimate the actual significance level. They also specify the value of the mean under the null hypothesis, H0. Usually, these distributions will be identical. The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean is entered first.

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Group Distribution(s)|H1

Specify the distribution of this group under the alternative hypothesis, H1. This distribution is used in the simulation that determines the power. A fundamental quantity in a power analysis is the amount of variation among the group means. In fact, classical power analysis formulas, this variation is summarized as the standard deviation of the means.

The important point to realize is that you must pay particular attention to the values you give to the means of these distributions because they are fundamental to the interpretation of the simulation.

For convenience in specifying a range of values, the parameters of the distribution can be specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean, M1, is entered first.

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)

Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Equivalence Margin

Specify the largest difference for which means from different groups will be considered equal. When specifying group distributions, it is possible to end up with scenarios where some means are slightly different from each other, even though they are intended to be equivalent. This often happens when specifying distributions of different forms (e.g. normal and gamma) for different groups, where the means are intended to be the same. The parameters used to specify different distributions do not always result in means that are EXACTLY equal. This value lets you control how different means can be and still be considered equal.

This value is not used to specify the hypothesized mean differences of interest. The hypothesized differences are specified using the means (or parameters used to calculate means) for the null and alternative distributions.

This value should be MUCH smaller than the hypothesized mean differences.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for *M0* in the distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for *M1* in the distribution specifications given above. Although it can be used wherever you want, *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B, C)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values for each letter using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Reports Tab

The Reports tab contains settings about the format of the output.

Select Output – Numeric Reports

Show Various Reports & Plots

These options let you specify whether you want to generate the standard reports and plots.

Show Inc’s & 95% C.I.

Checking this option causes an additional line to be printed showing a 95% confidence interval for both the power and actual alpha and half the width of the confidence interval (the increment).

Select Output – Plots

Show Comparative Reports & Plots

These options let you specify whether you want to generate reports and plots that compare the test statistics that are available.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Random Numbers

Random Number Pool Size

This is the size of the pool of values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Example 1 – Power at Various Sample Sizes

An experiment is being designed to compare the responses of three different treatments to a control. Previous studies have shown that the standard deviation within a group is 3.0. Researchers want to detect a shift in the mean of at least 2.0. To accomplish this, they set the mean of the control group to zero and the three treatment means to 2.0. They want to investigate sample sizes of 5, 10, 15, and 20 subjects per group.

Their primary analysis will be a set of Dunnett multiple comparison tests. They set the FWER to 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons of Treatments vs. a Control (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Each Treatment vs. Control (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**5 10 15 20**
Group Sample Size Pattern	**Equal**
MC Procedure	**Dunnett**
Simulations	**2000**
Grps 1	**1**
Control Distribution | H0	**N(M0 S)**
Control Distribution | H1	**N(M0 S)**
Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M1 S)**
Minimum Difference	**0.1**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**2**
S	**3**
Report Tab	
All reports except Comparative	**checked**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Simulation Summary Report

```
Summary of Simulations of the Control Group and the 3 Treatment Groups
MC Procedure: Dunnett’s M.C. Test

                Group    Total
         Any-   Smpl.    Smpl.      All-  S.D. of  S.D. of
Sim.    Pairs    Size     Size     Pairs    Means     Data   Actual   Target
No.     Power       n        N     Power    Sm|H1    SD|H1     FWER     FWER     M0       M1       S
1       0.187     5.0       20     0.016      0.9      3.0    0.055    0.050    0.0      2.0     3.0
      (0.017)  [0.170   0.204]   (0.005)   [0.010   0.021]  (0.010)   [0.045  0.064]

2       0.358    10.0       40     0.045      0.9      3.0    0.050    0.050    0.0      2.0     3.0
      (0.021)  [0.337   0.379]   (0.009)   [0.036   0.054]  (0.010)   [0.040  0.059]

3       0.507    15.0       60     0.098      0.9      3.0    0.042    0.050    0.0      2.0     3.0
      (0.022)  [0.485   0.529]   (0.013)   [0.085   0.111]  (0.009)   [0.033  0.051]

4       0.623    20.0       80     0.162      0.9      3.0    0.045    0.050    0.0      2.0     3.0
      (0.021)  [0.602   0.644]   (0.016)   [0.145   0.178]  (0.009)   [0.035  0.054]

Pool Size: 10000. Simulations: 2000. Run Time: 26.45 seconds.

Summary of Simulations Report Definitions
H0: the null hypothesis that each treatment mean is equal to the control mean.
H1: the alternative hypothesis that at least one treatment mean differs from the control mean.
Pair: each comparison of a treatment mean and a control mean is a 'pair'.
All-Pairs Power: the estimated probability of detecting all unequal pairs.
Any-Pairs Power: the estimated probability of detecting at least one unequal pair.
n: the average of the group sample sizes.
N: the combined sample size of all groups.
Family-Wise Error Rate (FWER): the probability of detecting at least one equal pair assuming H0.
Target FWER: the user-specified FWE.
Actual FWER: the FWER estimated by the alpha simulation.
Sm|H1: the standard deviation of the group means under H1.
SD|H1: the pooled, within-group standard deviation under H1.
Second Row: provides the precision and aconfidence interval based on the size of the simulation for
Any-Pairs Power, All-Pairs Power, and FWER. The format is (Precision) [95% LCL and UCL Alpha].

Summary Statements
A one-way design with 3 treatment groups and one control group has an average group sample size
of 5.0 for a total sample size of 20. This design achieved an any-pair power of 0.187 and an
all-pair power of 0.0155 using the Dunnett's Test procedure for comparing each treatment mean
with the control mean. The target family-wise error rate was 0.050 and the actual family-wise
error rate was 0.055. The average within group standard deviation assuming the alternative
distribution is 3.0. These results are based on 2000 Monte Carlo samples from the null
distributions: N(M0 S); N(M0 S); N(M0 S); and N(M0 S) and the alternative distributions: N(M0
S); N(M1 S); N(M1 S); and N(M1 S). Other parameters used in the simulation were: M0 = 0.0, M1 =
2.0, and S = 3.0.
```

This report shows the estimated any-pairs power, all-pairs power, and FWER for each scenario. The second row shows three 95% confidence intervals in brackets: the first for the any-pairs power, the second for the all-pairs power, and the third for the FWER. Half the width of each confidence interval is given in parentheses as a fundamental measure of the precision of the simulation. As the number of simulations is increased, the width of the confidence intervals will decrease.

Any-Pairs Power

This is the probability of detecting any of the significant pairs. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

All-Pairs Power

This is the probability of detecting all of the significant pairs. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

Group Sample Size n

This is the average of the individual group sample sizes.

Total Sample Size N

This is the total sample size of the study.

S.D. of Means Sm|H1

This is the standard deviation of the hypothesized means of the alternative distributions. Under the null hypothesis, this value is zero. This value represents the magnitude of the difference among the means that is being tested. It is roughly equal to the average difference between the group means and the overall mean.

Note that the effect size is the ratio of Sm|H1 and SD|H1

S.D. of Data SD|H1

This is the within-group standard deviation calculated from samples from the alternative distributions.

Actual FWER

This is the value of FWER (family-wise error rate) estimated by the simulation using the H0 distributions. It should be compared with the Target FWER to determine if the test procedure is accurate.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the Actual FWER.

Target FWER

The target value of FWER.

M0

This is the value entered for M0, the group means under H0.

M1

This is the value entered for M1, the group means under H1.

S

This is the value entered for S, the standard deviation.

Error-Rate Summary for H0 Simulation

```
Error Rate Summary from H0 (Alpha) Simulation of 4 Groups
MC Procedure: Dunnett’s M.C. Test
                                 Prop.
                                (No. of
                  Mean          Type-1
         No. of  No. of   Prop.  Errors             Mean     Min     Max
Sim.     Equal   Type-1  Type-1    > 0)   Target   Pairs   Pairs   Pairs
No.      Pairs   Errors  Errors    FWER     FWER   Alpha   Alpha   Alpha
1            3    0.067   0.022   0.055    0.050   0.022   0.020   0.025
2            3    0.060   0.020   0.050    0.050   0.020   0.017   0.023
3            3    0.048   0.016   0.042    0.050   0.016   0.016   0.017
4            3    0.055   0.018   0.045    0.050   0.018   0.018   0.019
```

This report shows the results of the H0 simulation. This simulation uses the H0 settings for each group. Its main purpose is to provide an estimate of the FWER.

No. of Equal Pairs

Since under H0 all means are equal, this is the number of unique pairs of the groups. Thus, this is the number of treatment groups.

Mean No. of Type-1 Errors

This is the average number of type-1 errors (false detections) per set (family).

Prop. Type-1 Errors

This is the proportion of type-1 errors (false detections) among all tests that were conducted.

Prop. (No. of Type-1 Errors>0) FWER

This is the proportion of the H0 simulations in which at least one type-1 error occurred. This is called the family-wise error rate.

Target FWER

This is the target value of FWER that was set by the user.

Mean Pairs Alpha

Alpha is the probability of rejecting H0 when H0 is true. It is a characteristic of an individual test. This is the average alpha value over all of the tests in the family.

Min Pairs Alpha

This is the minimum of all of the individual comparison alphas.

Max Pairs Alpha

This is the maximum of all of the individual comparison alphas.

Error-Rate Summary for H1 Simulation

```
Error Rate Summary from H1 (Power) Simulation of 4 Groups
MC Procedure: Dunnett's M.C. Test
                                                (FDR)
                                  Prop.  Prop.  Prop.  Prop.
      No. of  Mean   Mean         Equal  Uneq. Detect. Undet.   All    Any
      Equal/ No. of No. of         that   that   that   that  Uneq.  Uneq.  Mean   Min    Max
Sim.  Uneq.  False  False  Prop.   were   were   were   were  Pairs  Pairs  Pairs  Pairs  Pairs
No.   Pairs   Pos.   Neg. Errors Detect. Undet. Equal  Uneq.  Power  Power Power  Power  Power
1       0/3   0.00   2.74  0.913  0.000  0.913  0.000  1.000  0.016  0.187 0.088  0.082  0.092
2       0/3   0.00   2.45  0.816  0.000  0.816  0.000  1.000  0.045  0.358 0.184  0.182  0.188
3       0/3   0.00   2.13  0.711  0.000  0.711  0.000  1.000  0.098  0.507 0.289  0.286  0.292
4       0/3   0.00   1.84  0.613  0.000  0.613  0.000  1.000  0.162  0.623 0.388  0.380  0.392
```

This report shows the results of the H1 simulation. This simulation uses the H1 settings for each group. Its main purpose is to provide an estimate of the power.

No. of Equal Pairs/Unequal Pairs

The first value is the number of pairs for which the control mean and the treatment mean were equal under H1. The second value is the number of pairs for which the means were different under H1.

Mean No. False Positives

This is the average number of equal pairs that were declared as being unequal by the testing procedure. A *false positive* is a type-1 (alpha) error.

Mean No. False Negatives

This is the average number of unequal pairs that were not declared as being unequal by the testing procedure. A *false negative* is a type-2 (beta) error.

Prop. Errors

This is the proportion of type-1 and type-2 errors.

Prop. Equal that were Detect.

This is the proportion of the equal pairs in the H1 simulations that were declared as unequal.

Prop. Uneq. that were Undet.

This is the proportion of the unequal pairs in the H1 simulations that were not declared as being unequal.

Prop. Detect. that were Equal (FDR)

This is the proportion of detected pairs in the H1 simulations that were actually equal. This is often called the *false discovery rate*.

Prop. Undet. that were Uneq.

This is the proportion of undetected pairs in the H1 simulations that were actually unequal.

All Uneq. Pairs Power

This is the probability of detecting all of the pairs that were different in the H1 simulation.

Any Uneq. Pairs Power

This is the probability of detecting any of the pairs that were different in the H1 simulation.

Mean, Min, and Max Pairs Power

These items give the average, the minimum, and the maximum of the individual comparison powers from the H1 simulation.

Detail Model Report

```
Detailed Model Report for Simulation No. 1
Target FWER = 0.050, M0 = 0.0, M1 = 2.0, S = 3.0
MC Procedure: Dunnett's M.C. Test

Hypo.             Group                    Group     Ave.    Simulation
Type     Groups   Labels       n/N          Mean     S.D.    Model
H0       1        Cntl        5/20           0.0      3.0    N(M0 S)
H0       2-4      B1-B3       5/20           0.0      3.0    N(M0 S)
H0       All                              Sm=0.0      3.0

H1       1        Cntl        5/20           0.0      3.0    N(M0 S)
H1       2-4      B1-B3       5/20           2.0      3.0    N(M1 S)
H1       All                              Sm=0.9      3.0

Detailed Model Report for Simulation No. 2
Hypo.             Group                    Group     Ave.    Simulation
Type     Groups   Labels       n/N          Mean     S.D.    Model
H0       1        Cntl       10/40           0.0      3.0    N(M0 S)
H0       2-4      B1-B3      10/40           0.0      3.0    N(M0 S)
H0       All                              Sm=0.0      3.0

H1       1        Cntl       10/40           0.0      3.0    N(M0 S)
H1       2-4      B1-B3      10/40           2.0      3.0    N(M1 S)
H1       All                              Sm=0.9      3.0

Detailed Model Report for Simulation No. 3
Hypo.             Group                    Group     Ave.    Simulation
Type     Groups   Labels       n/N          Mean     S.D.    Model
H0       1        Cntl       15/60           0.0      3.0    N(M0 S)
H0       2-4      B1-B3      15/60           0.0      3.1    N(M0 S)
H0       All                              Sm=0.0      3.0

H1       1        Cntl       15/60           0.0      2.9    N(M0 S)
H1       2-4      B1-B3      15/60           2.0      3.0    N(M1 S)
H1       All                              Sm=0.9      3.0

Detailed Model Report for Simulation No. 4
Hypo.             Group                    Group     Ave.    Simulation
Type     Groups   Labels       n/N          Mean     S.D.    Model
H0       1        Cntl       20/80           0.0      3.0    N(M0 S)
H0       2-4      B1-B3      20/80           0.0      3.0    N(M0 S)
H0       All                              Sm=0.0      3.0

H1       1        Cntl       20/80           0.0      3.0    N(M0 S)
H1       2-4      B1-B3      20/80           2.0      3.0    N(M1 S)
H1       All                              Sm=0.9      3.0
```

This report shows details of each row of the previous reports.

Hypo. Type

This indicates which simulation is being reported on each row. H0 represents the null simulation and H1 represents the alternative simulation.

Groups

Each group in the simulation is assigned a number. This item shows the arbitrary group number that was assigned.

Group Labels

These are the labels that were used in the individual alpha-level reports. Note that the control group is labeled ‘Cntl’.

n/N

n is the average sample size of the groups. N is the total sample size across all groups.

Group Mean

These are the means of the individual groups as specified for the H0 and H1 simulations.

Ave. S.D.

This is the average standard deviation of all groups reported on each line. Note that it is calculated from the simulated data.

Simulation Model

This is the distribution that was used to simulate data for the groups reported on each line.

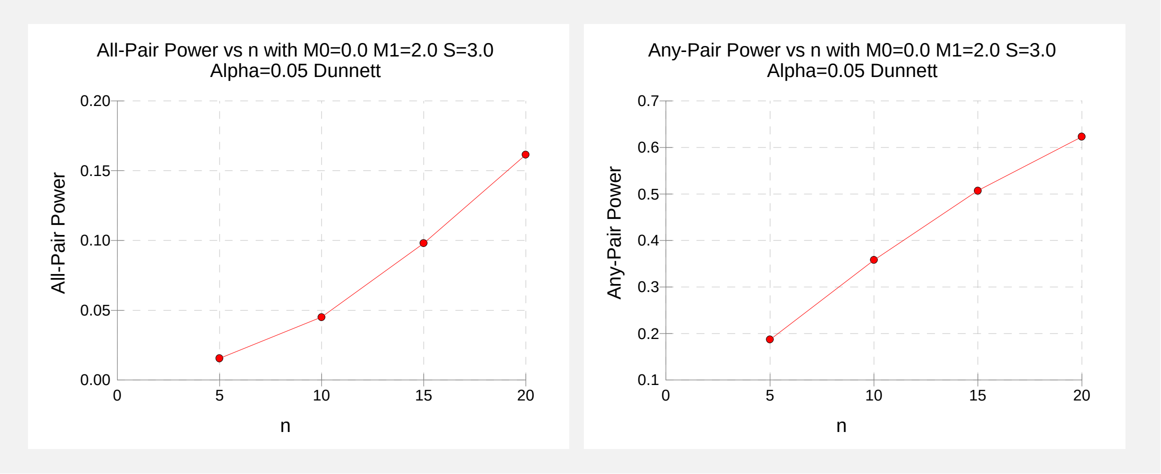
Probability of Rejecting Equality

```
Probability of Rejecting the Equality of Each Pair. Simulation No. 1

Group     Means      Cntl        B1        B2        B3
Cntl        0.0              0.082*    0.092*    0.090*
B1          2.0     0.025
B2          2.0     0.020
B3          2.0     0.022

Probability of Rejecting the Equality of Each Pair. Simulation No. 2

Group     Means      Cntl        B1        B2        B3
Cntl        0.0              0.188*    0.184*    0.182*
B1          2.0     0.023
B2          2.0     0.020
B3          2.0     0.017

Probability of Rejecting the Equality of Each Pair. Simulation No. 3

Group     Means      Cntl        B1        B2        B3
Cntl        0.0              0.292*    0.286*    0.289*
B1          2.0     0.016
B2          2.0     0.016
B3          2.0     0.017

Probability of Rejecting the Equality of Each Pair. Simulation No. 4

Group     Means      Cntl        B1        B2        B3
Cntl        0.0              0.392*    0.380*    0.392*
B1          2.0     0.019
B2          2.0     0.018
B3          2.0     0.018

Individual pairwise powers from the H1 (Power) simulation are shown in the upper-right section.
Individual pairwise significance levels from the H0 (Alpha) simulation are shown in the lower-left section.
* Starred values are the powers of pairs that are unequal under H1.
```

This report shows the individual probabilities of rejecting each pair. When a pair was actually different, the value is the power of that test. These power values are starred.

The results shown on the upper-right section of each simulation report are from the H1 simulation. The results shown on the lower-left section of the report are from the H0 simulation.

Plots Section



These plots give a visual presentation of the all-pairs power values and the any-pair power values.

Example 2 – Comparative Results

Continuing with Example 1, the researchers want to study the characteristics of alternative multiple comparison procedures.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons of Treatments vs. a Control (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Each Treatment vs. Control (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**5 10 15 20**
Group Sample Size Pattern	**Equal**
MC Procedure	**Dunnett**
Simulations	**2000**
Grps 1	**1**
Control Distribution | H0	**N(M0 S)**
Control Distribution | H1	**N(M0 S)**
Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M1 S)**

Data Tab (continued)

Minimum Difference **0.1**
M0 (Mean|H0) **0**
M1 (Mean|H1) **2**
S **3**

Reports Tab

Comparative Reports .... **Checked**
Comparative Any-Pair Power Plot .... **Checked**
Comparative All-Pair Power Plot............ **Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

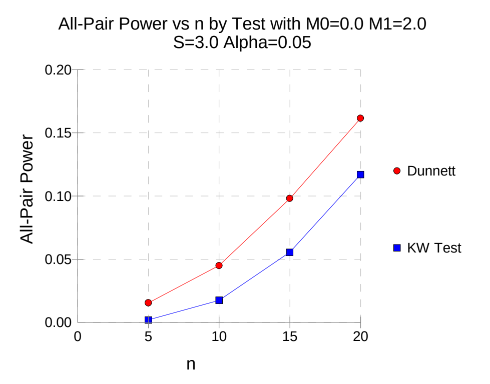
Power Comparison for Testing the Control Group Versus the 3 Treatment Groups

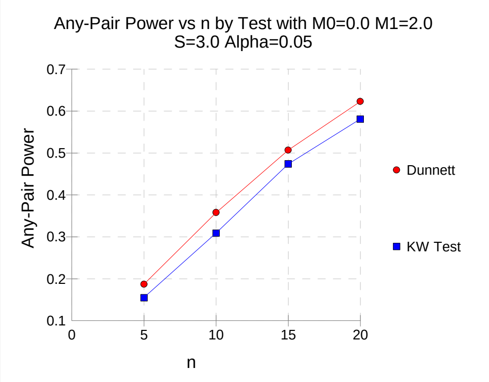
Sim. No.	Total Sample Size	Target Alpha	Dunnett All-Pair Power	Kruskal Wallis All-Pair Power	Dunnett All-Pair Power	Kruskal Wallis Any-Pair Power
1	20	0.050	0.016	0.002	0.187	0.155
2	40	0.050	0.045	0.018	0.358	0.309
3	60	0.050	0.098	0.056	0.507	0.474
4	80	0.050	0.162	0.117	0.623	0.581

Pool Size: 10000. Simulations: 2000. Run Time: 26.18 seconds.

Family-Wise FWER Comparison for Testing the Control Group Versus the 3 Treatment Groups

Sim. No.	Total Sample Size	Target FWER	Dunnett FWER	Kruskal Wallis FWER
1	20	0.050	0.055	0.044
2	40	0.050	0.050	0.040
3	60	0.050	0.042	0.034
4	80	0.050	0.045	0.038





These reports show the power and FWER of both of the multiple comparison procedures. In these simulations of groups from the normal distributions with equal variances, we see that the Dunnett’s procedure is the champion.

Example 3 – Validation using Dunnett

Murkerjee, Robertson, and Wright (1987) page 909 present an example of a sample size calculation which was first discussed on page 1116 of Dunnett (1955). In this example there are five treatments and one control. The control mean and four of the treatment means are the same, while the fifth treatment mean is different.

The value of the within-group standard deviation is 1.0. Four treatment means and the control mean are -0.182574 and the fifth treatment mean is 0.912871. The FWER is 0.05 in the article, but this is for a one-sided test. Since ***PASS*** is finding the power for two-sided tests, we set FWER at 0.10. When the per-group sample size is 16, the all-pairs power is 0.80. When the per-group sample size is 21, the all-pairs power is 0.90. Note that since there is only one treatment that is different from the control, the all-pairs power is equal to the any-pairs power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Comparisons of Treatments vs. a Control (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Each Treatment vs. Control (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.10**
n (Sample Size Multiplier)	**16 21**
Group Sample Size Pattern	**Equal**
MC Procedure	**Dunnett**
Simulations	**2000**
Grps 1	**1**
Control Distribution(s) | H0	**N(M0 S)**
Control Distribution(s) | H1	**N(M0 S)**
Grps 2	**1**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M1 S)**
Grps 3	**4**
Group 3 Distribution(s) | H0	**N(M0 S)**
Group 3 Distribution(s) | H1	**N(M0 S)**
Minimum Difference	**0.01**
M0 (Mean|H0)	**-0.182574**
M1 (Mean|H1)	**0.912871**
S	**1.0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Summary of Simulations of the Control Group and the 5 Treatment Groups
MC Procedure: Dunnett’s M.C. Test

                  Group     Total
          Any-    Smpl.     Smpl.      All-   S.D. of   S.D. of
Sim.     Pairs     Size      Size     Pairs     Means      Data    Actual    Target
No.      Power        n         N     Power    Sm|H1     SD|H1      FWER      FWER       M0        M1        S
1        0.795     16.0        96     0.795       0.4       1.0     0.095     0.100     -0.2       0.9      1.0
       (0.018)   [0.777    0.812]   (0.018)    [0.777    0.812]   (0.013)   [0.082   0.107]

2        0.894     21.0       126     0.894       0.4       1.0     0.112     0.100     -0.2       0.9      1.0
       (0.014)   [0.880    0.907]   (0.014)    [0.880    0.907]   (0.014)   [0.098   0.125]

Pool Size: 10000. Simulations: 1000. Run Time: 4.28 seconds.
```

For the first case when n = 16, ***PASS*** obtained a power of 0.795 which is very close to the value of 0.80 found by Muterjee et al. (1987). Indeed, 0.80 is within the confidence limits of 0.777 to 0.812. Similarly, the power for the second case when n = 21 is found as 0.894 which is very close to the article’s value of 0.90.

Chapter 590

Multiple Contrasts (Simulation)

Introduction

This procedure uses simulation to analyze the power and significance level of two multiple-comparison procedures that perform two-sided hypothesis tests of contrasts of the group means. These are the Dunn-Bonferroni test and the Dunn-Welch test. For each scenario, two simulations are run: one estimates the significance level and the other estimates the power.

The term *contrast* refers to a user-defined comparison of the group means. The term *multiple contrasts* refers to a set of such comparisons. An additional restriction imposed is that the contrast coefficients to sum to zero.

When several contrasts are tested, the interpretation of the results is more complex because of the problem of *multiplicity*. *Multiplicity* here refers to the fact that the probability of making at least one incorrect decision increases as the number of statistical tests increases. Methods for testing *multiple contrasts* have been developed to account for this multiplicity.

Error Rates

When dealing with several simultaneous statistical tests, both individual-wise and experiment wise error rates should be considered.

1. **Comparison-wise error rate**. This is the probability of a type-I error (rejecting a true H0) for a particular test. In the case of the five-group design, there are ten possible comparison-wise error rates, one for each of the ten possible pairs. We will denote this error rate α_c.

2. **Experiment-wise (or family-wise) error rate**. This is the probability of making one or more type-I errors in the set (family) of comparisons. We will denote this error rate α_f.

The relationship between these two error rates when the tests are independent is given by

$$\alpha_f = 1 - (1 - \alpha_c)^C$$

where C is the total number of contrasts. For example, if α_c is 0.05 and C is 10, α_f is 0.401.

There is about a 40% chance that at least one of the ten contrasts will be concluded to be non-zero when in fact they are not. When the tests are correlated, as they might be among a set of contrasts, the above formula provides an upper bound to the family-wise error rate.

The techniques described below provide control for α_f rather than α_c.

Technical Details

The One-Way Analysis of Variance Design

The discussion that follows is based on the common one-way analysis of variance design which may be summarized as follows. Suppose the responses Y_{ij} in k groups each follow a normal distribution with means $\mu_1, \mu_2, \cdots, \mu_k$ and unknown variance σ^2. Let $n_1, n_2, \cdots, n_k$ denote the number of subjects in each group. The control group is assumed to be group one.

The analysis of these responses is based on the sample means

$$\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}$$

and the pooled sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i \right)^2}{\sum_{i=1}^{k} \left(n_i - 1 \right)}$$

The *F* test is the usual method of analysis of the data from such a design, testing whether all of the means are equal. However, a significant *F* test does not indicate which of the groups are different, only that at least one is different. The analyst is left with the problem of determining which of the groups are different and by how much.

To determine the mean differences that are most importance, the researcher may specify a set of contrasts. These contrasts, called *a priori*, or, *planned*, contrasts should be specified before the experimental results are viewed.

The Dunn-Bonferroni procedure and the Dunn-Welch procedure have been developed to test these planned contrasts. The calculations associated with each of these tests are given below.

Contrasts

A contrast of the means is a stated difference among the means. The difference is constructed so that it represents a useful hypothesis. For example, suppose there are four groups, the first of which is a control group. It might be of interest to determine which treatments are statistically different from the control. That is, the differences $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, and $\mu_4 - \mu_1$ would be tested to determine if they are non-zero.

Contrasts are often simple differences between two means. However, they may involve more than just two means. For example, suppose the first two groups receive one treatment and the second two groups receive another treatment. The contrast (difference) that would be tested is $\left(\mu_1 + \mu_2 \right) - \left(\mu_3 + \mu_4 \right)$.

Every contrast can be represented by the list of contrast coefficients: the values by which the means are multiplied. Here are some examples of contrasts that might be of interest when the experiment involves four groups.

Difference	Coefficients
$\mu_2 - \mu_1$	-1, 1, 0, 0
$\mu_3 - \mu_1$	-1, 0, 1, 0
$(\mu_1 + \mu_2) - (\mu_3 + \mu_4)$	1, 1, -1, -1
$(\mu_1 + \mu_1 + \mu_1) - (\mu_2 + \mu_3 + \mu_4)$	3, -1, -1, -1
$(\mu_4 + \mu_4) - (\mu_2 + \mu_3)$	0, -1, -1, 2

Note that in each case, the coefficients sum to zero. This makes it possible to test whether the quantity is different from zero.

A lot is written about *orthogonal contrasts* which have the property that the sum of the products of corresponding coefficients is zero. For example, the sum of the products of the last two contrasts given above is 0(3) + (-1)(-1) + (-1)(-1) + (2)(-1) = 0 + 1 + 1 – 2 = 0, so these two contrasts are orthogonal. However, the first two contrasts are not orthogonal since (-1)(-1) + (1)(0) + (0)(1) + (0)(0) = 1 + 0 + 0 + 0 = 1 (not zero). Orthogonal contrasts have nice properties when the sample sizes are equal. Unfortunately, they lose those properties when the group sample sizes are unequal or when the data are not normally distributed.

The procedures described in this chapter do not require that the contrasts be orthogonal. Instead, you should focus on defining a set of contrasts that answer the research questions of interest.

Dunn-Bonferroni Test

Dunn (1964) developed a procedure for simultaneously testing several contrasts. This method is also discussed in Kirk (1982) pages 106 to 109. The method consists of testing each contrast with Student's *t* distribution with degrees of freedom equal to *N*-*k* with a Bonferroni adjustment of the alpha value. That is, the alpha value is divided by *C*, the number of contrasts simultaneously tested.

The test rejects H0 if

$$\frac{\left|\sum_{i=1}^{k} c_i \overline{Y}_i\right|}{\sqrt{\hat{\sigma}^2\left(\sum_{i=1}^{k}\frac{c_i^2}{n_i}\right)}} \geq \left|t_{1-\alpha/(2C),N-k}\right|$$

Note that this is a two-sided test of the hypothesis that $\sum_{i=1}^{k} c_i \mu_i = 0$ where $\sum_{i=1}^{k} c_i = 0$.

Dunn-Welch Test

Dunn (1964) developed a procedure for simultaneously testing several contrasts. This method, using Welch's (1947) modification for the unequal variances, is discussed in Kirk (1982) pages 100, 101, 106 - 109. The method consists of testing each contrast with Student's *t* distribution with degrees of freedom given below with a Bonferroni adjustment of the alpha value. That is, the alpha value is divided by *C*, the number of contrasts simultaneously tested.

The two-sided test statistic rejects H0 if

$$\frac{\left|\sum_{i=1}^{k} c_i \overline{Y}_i\right|}{\sqrt{\sum_{i=1}^{k} \frac{c_i^2 \hat{\sigma}_i^2}{n_i}}} \geq \left|t_{1-\alpha/(2C),v'}\right|$$

where

$$v' = \frac{\left(\sum_{i=1}^{k} \frac{c_i^2 \hat{\sigma}_i^2}{n_i}\right)^2}{\sum_{i=1}^{k} \frac{c_i^4 \hat{\sigma}_i^4}{n_i^2 \left(n_i - 1\right)}}$$

Definition of Power for Multiple Contrasts

The notion of power is well-defined for individual tests. Power is the probability of rejecting a false null hypothesis. However, this definition does not extend directly when there are a number of simultaneous tests. The two definitions that we emphasize in ***PASS*** where recommended by Ramsey (1978). They are *any-contrast power* and *all-contrasts power*. Other design characteristics, such as *average-contrast power* and *false-discovery rate*, are important to consider. However, our review of the statistical literature resulted in our focus on these two definitions of power.

Any-Contrast Power

Any-contrast power is the probability of detecting at least one of the contrasts that are actually non-zero.

All-Contrasts Power

All-contrast power is the probability of detecting all of the contrasts that are actually non-zero.

Simulation Details

Computer simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable. Computer simulation was once limited to mainframe computers. But, in recent years, as computer speeds have increased, simulation studies can be completed on desktop and laptop computers in a reasonable period of time.

The steps to a simulation study are

1. Specify how each test is to be carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.
2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the power of each test.

3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. The number rejected is used to calculate the significance level of each test.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in ***PASS*** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draws the random numbers from this pool. This second method can cut the running time of the simulation by 70%!

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data, Reports, and Options tabs. To find out more about using the other tabs such as Axes/Legend, Plot Text, or Template, go to the Procedure Window chapter.

Data 1 Tab

The Data 1 tab contains most of the parameters and options that you will be concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for: power or sample size (n). If you choose to solve for n, you must choose the type of power you want to solve for: any-contrast power or all-contrasts power. The value of the option *Power* will then represent this type of power.

Any-contrast power is the probability of detecting at least one of the non-zero contrasts. *All-contrast power* is the probability of detecting all non-zero contrasts.

Note that the search for n may take several minutes because a separate simulation must be run for each trial value of n. You may find it quicker and more informative to solve for the Power for a range of sample sizes.

Error Rates

Power or Beta

This option is only used when *Find (Solve For)* is set to *n (All-Contrast)* or *n (Any-Contrast)*.

Power is defined differently with multiple contrasts. Although many definitions are possible, two are adopted here. *Any-contrast power* is the probability of detecting at least one non-zero contrast. *All-contrasts power* is the probability of detecting all non-zero contrasts. As the number of contrasts is increased, these power probabilities will decrease because more tests are being conducted.

Since this is a probability, the range is between 0 and 1. Most researchers would like to have the power at least at 0.8. However, this may require extremely large sample sizes when the number of tests is large.

FWER (Alpha)

This option specifies one or more values of the *family-wise error rate* (FWER) which is the analog of alpha for multiple contrasts. FWER is the probability of falsely detecting (concluding that the means are different) at least one comparison for which the true means are the same. For independent tests, the relationship between the individual-comparison error rate (ICER) and FWER is given by the formulas

$$FWER = 1 - (1 - ICER)\verb|^|C$$

or

$$ICER = 1 - (1 - FWER)\verb|^|(1/C)$$

where '^' represents exponentiation (as in 4^2 = 16) and C represents the number of comparisons. For example, if C = 5 and FWER = 0.05, then ICER = 0.0102. Thus, the individual comparison tests must be conducted using a Type-1 error rate of 0.0102, which is much lower than the family-wise rate of 0.05.

The popular value for FWER remains at 0.05. However, if you have a large number of comparisons, you might decide that a larger value, such as 0.10, is appropriate.

Sample Size

n (Sample Size Multiplier)

This is the base, per group, sample size. One or more values separated by blanks or commas may be entered. A separate analysis is performed for each value listed here.

The group samples sizes are determined by multiplying this number by each of the Group Sample Size Pattern numbers. If the Group Sample Size Pattern numbers are represented by *m1*, *m2*, *m3*, …, *mk* and this value is represented by *n*, the group sample sizes *N1*, *N2*, *N3*, ..., *Nk* are calculated as follows:

N1=[n(m1)]

N2=[n(m2)]

N3=[n(m3)]

etc.

where the operator, [*X*] means the next integer after *X*, e.g. [3.1]=4.

For example, suppose there are three groups and the Group Sample Size Pattern is set to *1,2,3*. If n is 5, the resulting sample sizes will be 5, 10, and 15. If n is 50, the resulting group sample sizes will be 50, 100, and 150. If n is set to *2,4,6,8,10*, five sets of group sample sizes will be generated and an analysis run for each. These sets are:

2	4	6
4	8	12
6	12	18
8	16	24
10	20	30

As a second example, suppose there are three groups and the Group Sample Size Pattern is *0.2,0.3,0.5*. When the fractional Pattern values sum to one, n can be interpreted as the total sample size of all groups and the Pattern values as the proportion of the total in each group.

If n is 10, the three group sample sizes would be 2, 3, and 5.

If n is 20, the three group sample sizes would be 4, 6, and 10.

If n is 12, the three group sample sizes would be

(0.2)12 = 2.4 which is rounded up to the next whole integer, 3.

(0.3)12 = 3.6 which is rounded up to the next whole integer, 4.

(0.5)12 = 6.

Note that in this case, 3+4+6 does not equal n (which is 12). This can happen because of rounding.

Group Sample Size Pattern

The purpose of the group sample size pattern is to allow several groups with the same sample size to be generated without having to type each individually.

A set of positive, numeric values (one for each row of distributions) is entered here. Each item specified in this list applies to the whole row of distributions. For example, suppose the entry is *1 2 1* and Grps 1 = 3, Grps 2 = 1, Grps 3 = 2. The sample size pattern used would be *1 1 1 2 1 1*.

The sample size of group *i* is found by multiplying the ith number from this list by the value of *n* and rounding up to the next whole number. The number of values must match the number of groups, *g*. When too few numbers are entered, 1's are added. When too many numbers are entered, the extras are ignored.

- **Equal**

  If all sample sizes are to be equal, enter *Equal* here and the desired sample size in *n*. A set of *g* 1's will be used. This will result in *n1* = *n2* = … = *ng* = *n*. That is, all sample sizes are equal to *n*.

Test

MC Procedure

Specify which multiple contrast procedure is to be reported from the simulations. The choices are

- **Dunn-Bonferroni Test**

  This is the most popular and most often recommended.

- **Dunn-Welch Test**

  This is recommended when the group variances are very different.

Simulations

Simulations

This option specifies the number of iterations, *M*, used in the simulation. As the number of iterations is increased, the running time and accuracy are increased as well.

The precision of the simulated power estimates are calculated using the binomial distribution. Thus, confidence intervals may be constructed for various power values. The following table gives an estimate of the precision that is achieved for various simulation sizes when the power is either 0.50 or 0.95. The table values are interpreted as follows: a 95% confidence interval of the true power is given by the power reported by the simulation plus and minus the 'Precision' amount given in the table.

Simulation	Precision	Precision
Size	when	when
M	Power = 0.50	Power = 0.95
100	0.100	0.044
500	0.045	0.019
1000	0.032	0.014
2000	0.022	0.010
5000	0.014	0.006
10000	0.010	0.004
50000	0.004	0.002
100000	0.003	0.001

Notice that a simulation size of 1000 gives a precision of plus or minus 0.01 when the true power is 0.95. Also note that as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved.

Effect Size

These options specify the distributions to be used in the two simulations. The first option specifies the number of groups represented by the two distributions that follow. The second option specifies the distribution to be used in simulating the null hypothesis to determine the significance level (alpha). The third option specifies the distribution to be used in simulating the alternative hypothesis to determine the power.

Grps [1 – 3] (Grps 4 – 9 are found on the Data 2 tab)

This value specifies the number of groups specified by the H0 and H1 distribution statements to the right. Usually, you will enter '1' to specify a single H0 and a single H1 distribution, or you will enter '0' to indicate that the distributions specified on this line are to be ignored. This option lets you easily specify many identical distributions with a single phrase.

The total number of groups *g* is equal to the sum of the values for the three rows of distributions shown under the Data1 tab and the six rows of distributions shown under the Data2 tab.

Note that each item specified in the *Group Sample Size Pattern* option applies to the whole row of entries here. For example, suppose the *Group Sample Size Pattern* was *1 2 1* and Grps 1 = 3, Grps 2 = 1, and Grps 3 = 2. The sample size pattern would be *1 1 1 2 1 1*.

Group Distribution(s)|H0

This entry specifies the distribution of one or more groups under the null hypothesis, H0. The magnitude of the differences of the means of these distributions, which is often summarized as the standard deviation of the means, represents the magnitude of the mean differences specified under H0. Usually, the means are assumed to be equal under H0, so their standard deviation should be zero except for rounding.

These distributions are used in the simulations that estimate the actual significance level. They also specify the value of the mean under the null hypothesis, H0. Usually, these distributions will be identical. The parameters of each distribution are specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M0* is reserved for the value of the mean under the null hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean is entered first.

Beta=A(M0,A,B,Minimum)
Binomial=B(M0,N)
Cauchy=C(M0,Scale)
Constant=K(Value)
Exponential=E(M0)
F=F(M0,DF1)
Gamma=G(M0,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M0,SD)
Poisson=P(M0)
Student's T=T(M0,D)
Tukey's Lambda=L(M0,S,Skewness,Elongation)
Uniform=U(M0,Minimum)
Weibull=W(M0,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Finding the Value of the Mean of a Specified Distribution

Except for the multinomial distribution, the distributions have been parameterized in terms of their means since this is the parameter being tested. The mean of a distribution created as a linear combination of other distributions is found by applying the linear combination to the individual means. However, the mean of a distribution created by multiplying or dividing other distributions is not necessarily equal to applying the same function to the individual means. For example, the mean of 4N(4, 5) + 2N(5, 6) is 4*4 + 2*5 = 26, but the mean of 4N(4, 5) * 2N(5, 6) is not exactly 4*4*2*5 = 160 (although it is close).

Group Distribution(s)|H1

Specify the distribution of this group under the alternative hypothesis, H1. This distribution is used in the simulation that determines the power. A fundamental quantity in a power analysis is the amount of variation among the group means. In fact, classical power analysis formulas, this variation is summarized as the standard deviation of the means.

The important point to realize is that you must pay particular attention to the values you give to the means of these distributions because they are fundamental to the interpretation of the simulation.

For convenience in specifying a range of values, the parameters of the distribution can be specified using numbers or letters. If letters are used, their values are specified in the boxes below. The value *M1* is reserved for the value of the mean under the alternative hypothesis.

Following is a list of the distributions that are available and the syntax used to specify them. Note that, except for the multinomial, the distributions are parameterized so that the mean, M1, is entered first.

Beta=A(M1,A,B,Minimum)
Binomial=B(M1,N)
Cauchy=C(M1,Scale)
Constant=K(Value)
Exponential=E(M1)
F=F(M1,DF1)
Gamma=G(M1,A)
Multinomial=M(P1,P2,…,Pk)
Normal=N(M1,SD)
Poisson=P(M1)
Student's T=T(M1,D)
Tukey's Lambda=L(M1,S,Skewness,Elongation)
Uniform=U(M1,Minimum)
Weibull=W(M1,B)

Details of writing mixture distributions, combined distributions, and compound distributions are found in the chapter on Data Simulation and will not be repeated here.

Equivalence Margin

Specify the largest difference for which means from different groups will be considered equal. When specifying group distributions, it is possible to end up with scenarios where some means are slightly different from each other, even though they are intended to be equivalent. This often happens when specifying distributions of different forms (e.g. normal and gamma) for different groups, where the means are intended to be the same. The parameters used to specify different distributions do not always result in means that are EXACTLY equal. This value lets you control how different means can be and still be considered equal.

This value is not used to specify the hypothesized mean differences of interest. The hypothesized differences are specified using the means (or parameters used to calculate means) for the null and alternative distributions.

This value should be MUCH smaller than the hypothesized mean differences.

Effect Size – Distribution Parameters

M0 (Mean|H0)

These values are substituted for *M0* in the distribution specifications given above. *M0* is intended to be the value of the mean hypothesized by the null hypothesis, H0.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

M1 (Mean|H1)

These values are substituted for *M1* in the distribution specifications given above. Although it can be used wherever you want, *M1* is intended to be the value of the mean hypothesized by the alternative hypothesis, H1.

You can enter a list of values using the syntax *0 1 2 3* or *0 to 3 by 1*.

Parameter Values (S, A, B, C)

Enter the numeric value(s) of the parameters listed above. These values are substituted for the corresponding letter in all four distribution specifications.

You can enter a list of values for each letter using the syntax *0 1 2 3* or *0 to 3 by 1*.

You can also change the letter that is used as the name of this parameter using the pull-down menu to the side.

Contrasts Tab

Contrasts

Contrasts

These options specify the contrasts. You can specify as many contrasts as are necessary, but a penalty is paid in terms of reduced power for each additional contrast. Thus, the number of contrasts should be limited to those that are most important to the study.

A contrast is a weighted average of the *k* (*k* = number of groups) group means in which the weights (coefficients) sum to zero. Each successive coefficient is applied to the corresponding group mean. For example, suppose *k* = 3 and the first group is a control group. Two contrasts that might be of interest are *-1 1 0* and *-1 0 1*. These are interpreted as *(-1)Mean1 + (1)Mean2 + (0)Mean3* and *(-1)Mean1 + (0)Mean2 + (1)Mean3*, respectively. Notice that the coefficients in each set sum to zero.

Several predefined sets of contrasts are available or you can specify your own. There is no set number of contrasts that must (or may) be specified, but fewer contrasts result in higher power and smaller required samples sizes.

Possible entries are given next.

- **Individual Contrasts**

  Enter a set of numbers, separated by blanks. One coefficient must be entered for each group with one set per box. Examples of valid contrasts are

  -1 1
  -1 0 1
  0 1 -2 1
  -4 1 1 2

- **Each With First**

  This option generates *k*-1 contrasts appropriate for comparing each of the remaining groups with the first group. This might be used when the first group is a control group. If *k* = 4, the 3 contrasts are

  -1 1 0 0
  -1 0 1 0
  -1 0 0 1

- **Each With Last**

  This option generates *k*-1 contrasts appropriate for comparing each of the first *k*-1 groups with the last group. This might be used when the last group is a control group. If *k* = 4, the 3 contrasts are

  -1 0 0 1
  0 -1 0 1
  0 0 -1 1

- **Each With Next**

  This option generates *k*-1 contrasts appropriate for comparing each group with the next group. If *k* = 4, the 3 contrasts are

  -1 1 0 0
  0 -1 1 0
  0 0 -1 1

- **Each With Remaining**

  Each group mean is compared with the average of those remaining to the right. Suppose *k*=4, the 3 contrasts are

  -3 1 1 1
  0 -2 1 1
  0 0 -1 1

- **Each With All Others**

  Each group mean is compared with the average of the other groups. Suppose *k*=4, the 4 contrasts are

  -3 1 1 1
  1 -3 1 1
  1 1 -3 1
  1 1 1 -3

- **Progressive Split**

  The first groups are compared to the last groups. The dividing point moves from left to right. Suppose *k*=5, the 4 contrasts are

  -4 1 1 1 1
  -3 -3 2 2 2
  -2 -2 -2 3 3
  -1 -1 -1 -1 4.

Reports Tab

The Reports tab contains settings about the format of the output.

Select Output – Numeric Reports

Show Various Reports & Plots

These options let you specify whether you want to generate the standard reports and plots.

Show Inc's & 95% C.I.

Checking this option causes an additional line to be printed showing a 95% confidence interval for both the power and actual alpha and half the width of the confidence interval (the increment).

Select Output – Plots

Show Comparative Reports & Plots

These options let you specify whether you want to generate reports and plots that compare the test statistics that are available.

Options Tab

The Options tab contains limits on the number of iterations and various options about individual tests.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the sample size is aborted. When the maximum number of iterations is reached without convergence, the sample size is left blank. We recommend a value of at least 500.

Random Numbers

Random Number Pool Size

This is the size of the pool of values from which the random samples will be drawn. Pools should be at least the maximum of 10,000 and twice the number of simulations. You can enter *Automatic* and an appropriate value will be calculated.

If you do not want to draw numbers from a pool, enter 0 here.

Example 1 – Power at Various Sample Sizes

A study is being planned to find the threshold level of a certain drug. Below this threshold level, the response has little change. Once the threshold level is reached, there is a sizeable jump in the mean response rate. Little change in the response occurs as the drug level is increased above the threshold. Scientists believe that the threshold level is between 3 and 7—their best estimate, based on previous studies, is 5. Previous studies have shown that the standard deviation within a group is 3.0.

In order to find the threshold, they design a study with five levels: 3.0, 4.0, 5.0, 6.0, and 7.0. Since there is no trend in the mean value (only a sudden shift) as the dose level is increased, they decide to test the following hypotheses:

Difference	Coefficients
$\mu_2 - \mu_1$	-1, 1, 0, 0, 0
$\mu_3 - \mu_2$	0, -1, 1, 0, 0
$\mu_4 - \mu_3$	0, 0, -1, 1, 0
$\mu_5 - \mu_4$	0, 0, 0, -1, 1

Notice that this set of hypotheses answers the question directly. An overall F-test would test the hypothesis that at least one mean is different, but it would not indicate which is different. The question might be settled by considering all possible pairs, but there are ten pairs, so ten hypothesis tests would have to be considered instead of only four—decreasing the power.

Researchers want to detect a shift in the mean as small as 2.0. Hence, they want to study the power when the means are 0.0, 0.0, 2.0, 2.0, 2.0. They want to investigate sample sizes of 10, 30, 50, and 70 subjects per group.

They have no reason to assume that the variance will change a great deal from group to group, so they decide to analyze the data using the Dunn-Bonferroni procedure. They set the FWER to 0.05. Note that, based on these means, only the second of the four contrasts will be significant, so the any-contrast power will be the same as the all-contrast power.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Contrasts (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Contrasts (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**10 30 50 70**
Group Sample Size Pattern	**Equal**
MC Procedure	**Dunn-Bonferroni**
Simulations	**2000**

Data Tab (continued)

Grps 1..**2**
Control Distribution | H0..........................**N(M0 S)**
Control Distribution | H1..........................**N(M0 S)**
Grps 2..**3**
Group 2 Distribution(s) | H0 ....**N(M0 S)**
Group 2 Distribution(s) | H1 ....**N(M1 S)**
Minimum Difference **0.1**
M0 (Mean|H0) **0**
M1 (Mean|H1) **2**
S **3**

Contrasts Tab

Contrasts ....**Each With Next**

Reports Tab

All reports except Comparative..............**checked**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Simulation Summary Report

```
Summary of Simulations for Testing Multiple Contrasts of 5 Groups
MC Procedure: Dunn-Bonferroni Test

                 Group     Total
        Any-     Smpl.     Smpl.       All-    S.D. of   S.D. of
Sim.   Cont.      Size      Size      Cont.      Means      Data    Actual    Target
No.    Power         n         N      Power      Sm|H1     SD|H1      FWER      FWER      M0      M1      S
1      0.136      10.0        50      0.136        1.0       3.0     0.048     0.050     0.0     2.0    3.0
     (0.015)    [0.121    0.151]    (0.015)     [0.121    0.151]   (0.009)    [0.038   0.057]

2      0.509      30.0       150      0.509        1.0       3.0     0.038     0.050     0.0     2.0    3.0
     (0.022)    [0.487    0.530]    (0.022)     [0.487    0.530]   (0.008)    [0.030   0.046]

3      0.796      50.0       250      0.796        1.0       3.0     0.048     0.050     0.0     2.0    3.0
     (0.018)    [0.778    0.813]    (0.018)     [0.778    0.813]   (0.009)    [0.038   0.057]

4      0.924      70.0       350      0.924        1.0       3.0     0.041     0.050     0.0     2.0    3.0
     (0.012)    [0.912    0.936]    (0.012)     [0.912    0.936]   (0.009)    [0.032   0.050]
Pool Size: 10000. Simulations: 2000. Run Time: 56.58 seconds.

Summary of Simulations Report Definitions
H0: the null hypothesis that the contrast of the means is zero.
H1: the alternative hypothesis that the contrast of the means is not zero.
Cont.: abbreviates 'Contrast'. Refers to a weighted average of the means whose weights sum to zero.
All-Cont. Power: the estimated probability of detecting all unequal contrasts.
Any-Cont. Power: the estimated probability of detecting at least one unequal contrasts.
n: the average of the group sample sizes.
N: the combined sample size of all groups.
Family-Wise Error Rate (FWER): the probability of detecting at least one zero contrast assuming H0.
Target FWER: the user-specified FWE.
Actual FWER: the FWER estimated by the alpha simulation.
Sm|H1: the standard deviation of the group means under H1.
SD|H1: the pooled, within-group standard deviation under H1.
Second Row: provides the precision and a confidence interval based on the size of the simulation for
Any-Contrast Power, All-Contrasts Power, and FWER. The format is (Precision) [95% LCL and UCL Alpha].
```

Summary Statements

A one-way design with 5 groups has an average group sample size of 10.0 for a total sample size of 50. This design achieved an any-contrast power of 0.136 and an all-contrast power of 0.136 using the Dunn-Bonferroni Test procedure for comparing each contrast of the group means with zero. The target family-wise error rate was 0.050 and the actual family-wise error rate was 0.048. The average within group standard deviation assuming the alternative distribution is 3.0. These results are based on 2000 Monte Carlo samples from the null distributions: N(M0 S); N(M0 S); N(M0 S); N(M0 S); and N(M0 S) and the alternative distributions: N(M0 S); N(M0 S); N(M1 S); N(M1 S); and N(M1 S). Other parameters used in the simulation were: M0 = 0.0, M1 = 2.0, and S = 3.0.

This report shows that a group sample size of about 50 will be needed to achieve 80% power or about 70 for 90% power.

Any-Cont. Power

This is the probability of detecting any of the significant contrasts. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

All- Cont. Power

This is the probability of detecting all of the significant contrasts. This value is estimated by the simulation using the H1 distributions.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the estimated power.

Group Sample Size n

This is the average of the individual group sample sizes.

Total Sample Size N

This is the total sample size of the study.

S.D. of Means Sm|H1

This is the standard deviation of the hypothesized means of the alternative distributions. Under the null hypothesis this value is zero. It represents the magnitude of the difference among the means. It is roughly equal to the average difference between the group means and the overall mean.

Note that the effect size is the ratio of Sm|H1 and SD|H1.

S.D. of Data SD|H1

This is the within-group standard deviation calculated from samples from the alternative distributions.

Actual FWER

This is the value of FWER (family-wise error rate) estimated by the simulation using the H0 distributions. It should be compared with the Target FWER to determine if the test procedure is accurate.

Note that a precision value (half the width of its confidence interval) and a confidence interval are shown on the line below this row. These values provide the precision of the Actual FWER.

Target FWER

This is the target value of FWER that was set by the user.

M0

This is the value entered for M0, the group means under H0.

M1

This is the value entered for M1, the group means under H1.

S

This is the value entered for S, the standard deviation.

Error-Rate Summary for H0 Simulation

```
Error Rate Summary from H0 (Alpha) Simulation of 5 Groups
MC Procedure: Dunn-Bonferroni Test
                                  Prop.
                                (No. of
                 Mean           Type-1
        No. of  No. of   Prop.  Errors            Mean     Min     Max
Sim.      Zero  Type-1  Type-1    > 0)  Target   Cont.   Cont.   Cont.
No.      Cont.  Errors  Errors    FWER    FWER   Alpha   Alpha   Alpha
1            4   0.053   0.013   0.048   0.050   0.013   0.010   0.016
2            4   0.042   0.010   0.038   0.050   0.010   0.009   0.012
3            4   0.056   0.014   0.048   0.050   0.014   0.012   0.016
4            4   0.047   0.012   0.041   0.050   0.012   0.010   0.014
```

This report shows the results of the H0 simulation. This simulation uses the H0 settings for each group. Its main purpose is to provide an estimate of the FWER.

No. of Zero Cont.

Since under H0 all means are equal, this is the number of contrasts.

Mean No. of Type-1 Errors

This is the average number of type-1 errors (false detections) per set (family).

Prop. Type-1 Errors

This is the proportion of type-1 errors (false detections) among all tests that were conducted.

Prop. (No. of Type-1 Errors>0) FWER

This is the proportion of the H0 simulations in which at least one type-1 error occurred. This is called the family-wise error rate.

Target FWER

This is the target value of FWER that was set by the user.

Mean Cont. Alpha

Alpha is the probability of rejecting H0 when H0 is true. It is a characteristic of an individual test. This is the average individual alpha value over all of the contrasts.

Min Cont. Alpha

This is the minimum of all contrast alphas.

Max Cont. Alpha

This is the maximum of all contrast alphas.

Error-Rate Summary for H1 Simulation

```
Error-Rate Summary from H1 (Power) Simulation of 5 Groups
MC Procedure: Dunn-Bonferroni Test
                                                        (FDR)
                                        Prop.   Prop.   Prop.   Prop.
       No. of    Mean    Mean            Zero  Non-0. Detect.  Undet.     All     Any
        Zero/  No. of  No. of            that    that    that    that   Non-0   Non-0    Mean     Min     Max
Sim.    Non-0   False   False   Prop.    were    were    were    were   Cont.   Cont.   Cont.   Cont.   Cont.
No.     Cont.    Pos.    Neg.  Errors  Detect.  Undet.    Zero   Non-0   Power   Power   Power   Power   Power
1         3/1    0.04    0.86   0.226   0.013   0.864   0.223   0.226   0.136   0.136   0.044   0.010   0.136
2         3/1    0.03    0.49   0.131   0.011   0.492   0.061   0.142   0.509   0.509   0.135   0.009   0.509
3         3/1    0.04    0.20   0.062   0.014   0.205   0.050   0.065   0.796   0.796   0.209   0.009   0.796
4         3/1    0.03    0.08   0.027   0.010   0.076   0.031   0.025   0.924   0.924   0.239   0.008   0.924
```

This report shows the results of the H1 simulation. This simulation uses the H1 settings for each group. Its main purpose is to provide an estimate of the power.

No. of Zero/Non-0 Cont.

The first value is the number of contrasts that were zero under H1. The second value is the number of contrasts that were non-zero under H1.

Mean No. False Positives

This is the average number of zero contrasts that were declared as being non-zero by the testing procedure. A *false positive* is a type-1 (alpha) error.

Mean No. False Negatives

This is the average number of non-zero contrasts that were not declared as being non-zero by the testing procedure. A *false negative* is a type-2 (beta) error.

Prop. Errors

This is the proportion of type-1 and type-2 errors.

Prop. Equal that were Detect.

This is the proportion of the zero contrasts in the H1 simulations that were declared as non-zero.

Prop. Uneq. that were Undet.

This is the proportion of non-zero contrasts in the H1 simulations that were not declared as being non-zero.

Prop. Detect. that were Zero (FDR)

This is the proportion of all detected contrasts in the H1 simulations that were actually zero. This is often called the *false discovery rate*.

Prop. Undet. that were Non-0.

This is the proportion of undetected contrasts in the H1 simulations that were actually non-zero.

All Non-0 Cont. Power

This is the probability of detecting all non-zero contrasts in the H1 simulation.

Any Non-0 Cont. Power

This is the probability of detecting any non-zero contrasts in the H1 simulation.

Mean, Min, and Max Cont. Power

These items give the average, the minimum, and the maximum of the contrast powers from the H1 simulation.

Detail Model Report

```
Detailed Model Report for Simulation No. 1
Target FWER = 0.050, M0 = 0.0, M1 = 2.0, S = 3.0
MC Procedure: Dunn-Bonferroni Test

Hypo.            Group                Group      Ave.   Simulation
Type    Groups   Labels      n/N      Mean       S.D.   Model
H0      1-2      A1-A2      10/50      0.0        3.1   N(M0 S)
H0      3-5      B1-B3      10/50      0.0        3.0   N(M0 S)
H0      All                         Sm=0.0        3.0

H1      1-2      A1-A2      10/50      0.0        3.0   N(M0 S)
H1      3-5      B1-B3      10/50      2.0        3.0   N(M1 S)
H1      All                         Sm=1.0        3.0

Detailed Model Report for Simulation No. 2
Hypo.            Group                Group      Ave.   Simulation
Type    Groups   Labels      n/N      Mean       S.D.   Model
H0      1-2      A1-A2     30/150      0.0        3.0   N(M0 S)
H0      3-5      B1-B3     30/150      0.0        3.0   N(M0 S)
H0      All                         Sm=0.0        3.0

H1      1-2      A1-A2     30/150      0.0        3.0   N(M0 S)
H1      3-5      B1-B3     30/150      2.0        3.0   N(M1 S)
H1      All                         Sm=1.0        3.0

Detailed Model Report for Simulation No. 3
Hypo.            Group                Group      Ave.   Simulation
Type    Groups   Labels      n/N      Mean       S.D.   Model
H0      1-2      A1-A2     30/150      0.0        3.0   N(M0 S)
H0      3-5      B1-B3     30/150      0.0        3.0   N(M0 S)
H0      All                         Sm=0.0        3.0

H1      1-2      A1-A2     30/150      0.0        3.0   N(M0 S)
H1      3-5      B1-B3     30/150      2.0        3.0   N(M1 S)
H1      All                         Sm=1.0        3.0

Detailed Model Report for Simulation No. 4
Hypo.            Group                Group      Ave.   Simulation
Type    Groups   Labels      n/N      Mean       S.D.   Model
H0      1-2      A1-A2     70/350      0.0        3.0   N(M0 S)
H0      3-5      B1-B3     70/350      0.0        3.0   N(M0 S)
H0      All                         Sm=0.0        3.0

H1      1-2      A1-A2     70/350      0.0        3.0   N(M0 S)
H1      3-5      B1-B3     70/350      2.0        3.0   N(M1 S)
H1      All                         Sm=1.0        3.0
```

This report shows details of each row of the previous reports.

Hypo. Type

This indicates which simulation is being reported on each row. H0 represents the null simulation and H1 represents the alternative simulation.

Groups

Each group in the simulation is assigned a number. This item shows the arbitrary group number that was assigned.

Group Labels

These are the labels that were used in the individual alpha-level reports.

n/N

n is the average sample size of the groups. N is the total sample size across all groups.

Group Mean

These are the means of the individual groups as specified for the H0 and H1 simulations.

Ave. S.D.

This is the average standard deviation of all groups reported on each line. Note that it is calculated from the simulated data.

Simulation Model

This is the distribution that was used to simulate data for the groups reported on each line.

List of Contrast Coefficients

```
List of Contrast Coefficients

               Groups
Contrasts         A1        A2        B1        B2        B3
Con1            -1.0       1.0       0.0       0.0       0.0
Con2             0.0      -1.0       1.0       0.0       0.0
Con3             0.0       0.0      -1.0       1.0       0.0
Con4             0.0       0.0       0.0      -1.0       1.0
The contrasts are shown down the rows. The groups are shown across the columns.
The coefficients (weights) are shown as the body of the table.
```

This report shows values of the contrast coefficients so you can double-check that they are what was intended.

Probability of Rejecting Individual Contrasts

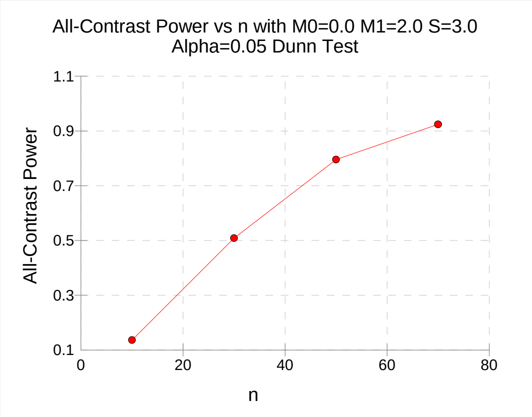
```
Probability of Rejecting Individual Contrasts.  Simulation No. 1

Contrasts  Alpha     Power
Con1       0.013     0.010
Con2       0.016     0.136
Con3       0.010     0.016
Con4       0.014     0.014
Alpha: probability of rejecting hypothesis that contrast is zero under alpha (H0) simulation.
Power: probability of rejecting hypothesis that contrast is zero under power (H1) simulation.

Probability of Rejecting Individual Contrasts.  Simulation No. 2

Contrasts  Alpha     Power
Con1       0.010     0.009
Con2       0.011     0.509
Con3       0.012     0.013
Con4       0.009     0.012
```

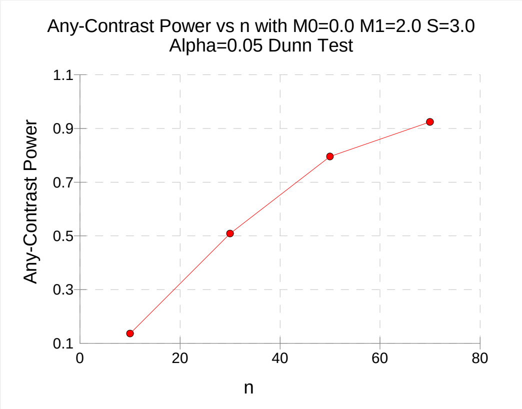
```
Probability of Rejecting Individual Contrasts.  Simulation No. 3

Contrasts  Alpha     Power
Con1        0.013     0.014
Con2        0.015     0.796
Con3        0.016     0.019
Con4        0.012     0.009

Probability of Rejecting Individual Contrasts.  Simulation No. 4

Contrasts  Alpha     Power
Con1        0.010     0.012
Con2        0.014     0.924
Con3        0.011     0.011
Con4        0.012     0.008
```

This report shows alpha and individual power for each contrast for each simulation that was run.

In this example, only the second contrast was non-zero, so that is the only one which has large values for the power.

Plots Section





These plots give a visual presentation of the all-contrasts power values and the any-contrast power values.

Example 2 – Comparative Results

Continuing with Example 1, the researchers want to study the characteristics of alternative multiple contrast procedures.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Contrasts (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Contrasts (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**10 30 50 70**
Data Tab (continued)	
Group Sample Size Pattern	**Equal**
MC Procedure	**Dunn-Bonferroni**
Simulations	**2000**
Grps 1	**2**
Control Distribution | H0	**N(M0 S)**
Control Distribution | H1	**N(M0 S)**
Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M1 S)**
Minimum Difference	**0.1**
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**2**
S	**3**
Contrasts Tab	
Contrasts	**Each With Next**
Reports Tab	
Comparative Reports	**Checked**
Comparative Any-Contrast Power Plot	**Checked**
Comparative All-Contrast Power Plot	**Checked**

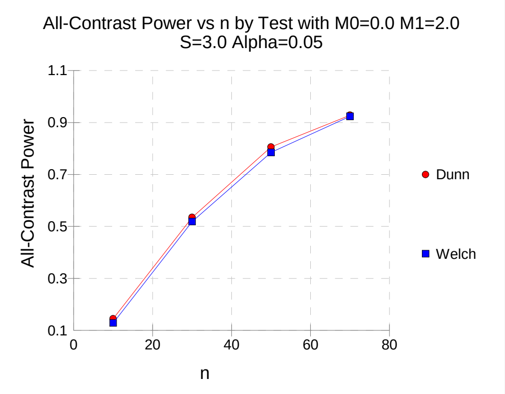
Output

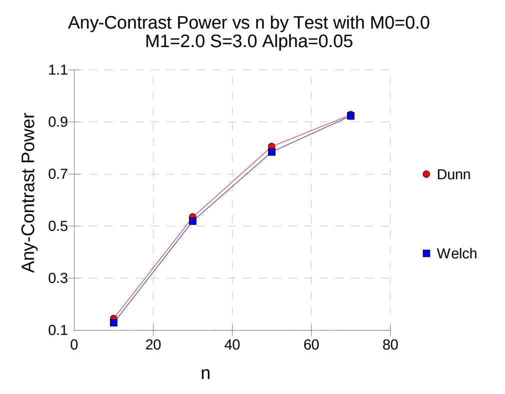
Click the Run button to perform the calculations and generate the following output.

Numeric Results and Plots

```
Power Comparison for Simultaneously Testing Multiple Contrasts of 5 Groups
                                        Dunn          Dunn          Dunn          Dunn
             Total                Bonferroni         Welch    Bonferroni         Welch
Sim.        Sample       Target    All-Cont.     All-Cont.     Any-Cont.     Any-Cont.
No.           Size        Alpha        Power         Power         Power         Power
1               50        0.050        0.145         0.129         0.145         0.129
2              150        0.050        0.535         0.519         0.535         0.519
3              250        0.050        0.806         0.785         0.806         0.785
4              350        0.050        0.928         0.924         0.928         0.924
Pool Size: 10000. Simulations: 2000. Run Time: 5.53 minutes.
```

```
Family-Wise Error-Rate Comparison for Simultaneously Testing Multiple Contrasts of 5 Groups

             Total                      Dunn          Dunn
Sim.        Sample       Target   Bonferroni         Welch
No.           Size         FWER         FWER          FWER
1               50        0.050        0.039         0.039
2              150        0.050        0.041         0.046
3              250        0.050        0.050         0.047
4              350        0.050        0.051         0.044
```


All-Contrast Power vs n by Test with M0=0.0 M1=2.0 S=3.0 Alpha=0.05


Any-Contrast Power vs n by Test with M0=0.0 M1=2.0 S=3.0 Alpha=0.05

These reports show the power and FWER of both multiple contrast procedures. In these simulations of groups from the normal distributions with equal variances, there is little difference in the power of the two procedures.

Example 3 – Validation

We could not find an article that gives power values for this test, so we decided to validate the procedure by comparing its results to those of the one-way ANOVA procedure which allows a single contrast to be tested. Using the settings of Example 1 and using the contrast ‘0, -1, 1, 0, 0’, we obtained the following powers: 0.3085, 0.7274, 0.9131, and 0.9758.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Multiple Contrasts (Simulation)** procedure window by clicking on **Means**, then **Multiple Comparisons**, then **Multiple Contrasts (Simulation)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power**
Power	*Ignored since this is the Find setting*
FWER (Alpha)	**0.05**
n (Sample Size Multiplier)	**10 30 50 70**
Group Sample Size Pattern	**Equal**
MC Procedure	**Dunn-Bonferroni**
Simulations	**2000**
Grps 1	**2**
Control Distribution | H0	**N(M0 S)**
Control Distribution | H1	**N(M0 S)**
Grps 2	**3**
Group 2 Distribution(s) | H0	**N(M0 S)**
Group 2 Distribution(s) | H1	**N(M1 S)**
Minimum Difference	**0.1**
Data Tab (continued)	
M0 (Mean|H0)	**0**
M1 (Mean|H1)	**2**
S	**3**
Contrasts Tab	
Contrasts	**0 -1 1 0 0**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Summary of Simulations for Testing Multiple Contrasts of 5 Groups
MC Procedure: Dunn-Bonferroni Test

                   Group     Total
          Any-     Smpl.     Smpl.      All-   S.D. of   S.D. of
Sim.     Cont.      Size      Size     Cont.     Means      Data    Actual    Target
No.      Power         n         N     Power     Sm|H1     SD|H1      FWER      FWER      M0      M1       S
1        0.297      10.0        50     0.297       1.0       3.0     0.050     0.050     0.0     2.0     3.0
       (0.020)    [0.276    0.317]   (0.020)    [0.276    0.317]   (0.010)    [0.040  0.059]

2        0.741      30.0       150     0.741       1.0       3.0     0.050     0.050     0.0     2.0     3.0
       (0.019)    [0.721    0.760]   (0.019)    [0.721    0.760]   (0.010)    [0.040  0.060]

3        0.914      50.0       250     0.914       1.0       3.0     0.061     0.050     0.0     2.0     3.0
       (0.012)    [0.901    0.926]   (0.012)    [0.901    0.926]   (0.010)    [0.050  0.071]

4        0.974      70.0       350     0.974       1.0       3.0     0.055     0.050     0.0     2.0     3.0
       (0.007)    [0.967    0.981]   (0.007)    [0.967    0.981]   (0.010)    [0.045  0.065]

Pool Size: 10000. Simulations: 2000. Run Time: 56.58 seconds.
```

In each case, the confidence interval includes the actual value. That is, 0.3085 is between 0.276 and 0.317, 0.7274 is between 0.721 and 0.760, 0.9131 is between 0.901 and 0.926, and 0.9758 is between 0.967 and 0.981. This validates the procedure.

590-26 Multiple Contrasts (Simulation)

Chapter 600

Hotelling’s T^2

Introduction

This module calculates power for Hotelling’s one-group, and two-group, T-squared (T2) test statistics. Hotelling’s T2 is an extension of the univariate t-tests in which the number of response variables is greater than one. In the two-group case, these results may also be obtained using ***PASS’s*** MANOVA test.

Assumptions

The following assumptions are made when using Hotelling’s T2 to analyze one or two groups of data.

1. The response variables are continuous.
2. The residuals follow the multivariate normal probability distribution with mean zero and constant variance-covariance matrix.
3. The subjects are independent.

Technical Details

The formulas used to perform a Hotelling’s T2 power analysis provide exact answers if the above assumptions are met. These formulas can be found in many places. We use the results in Rencher (1998). We refer you to that reference for more details.

One-Group Case

In this case, a set of *N* observations is available on *p* response variables. We assume that all *N* observations have the same multivariate normal distribution with mean vector μ and variance covariance matrix Σ and that Hotelling’s *T*2 is used for testing the null hypothesis that $\mu = \mu_0$ versus the alternative that $\mu = \mu_A$ where at least one component of μ_A is different from the corresponding component of μ_0. Usually, the vector μ_0 is a vector of zeros.

The value of *T*2 is computed using the formula

$$T^2_{p,N-1} = N\left(\bar{y} - \mu_0\right)' S^{-1}\left(\bar{y} - \mu_0\right)$$

where $\bar{y}$ is the vector of sample means and *S* is the sample variance-covariance matrix.

To calculate power we need the non-centrality parameter for this distribution. This non-centrality parameter is defined as follows

$$\begin{aligned}\lambda &= N(\mu_A - \mu_0)' \varSigma^{-1}(\mu_A - \mu_0) \\ &= N\varDelta^2\end{aligned}$$

where

$$\varDelta = \sqrt{(\mu_A - \mu_0)' \varSigma^{-1}(\mu_A - \mu_0)}$$

We define $\varDelta$ as *effect size* because it provides a expression for the magnitude of the standardized difference between the null and alternative means.

Using this non-centrality parameter, the power of the Hotelling's T2 may be calculated for any value of the means and standard deviations. Since there is a simple relationship between the non-central T2 and the non-central *F*, calculations are actually based on the non-central *F* using the formula

$$\beta = \Pr\left(F' < F'_{\alpha,df1,df2,\lambda}\right)$$

where

$$df1 = p$$
$$df2 = N - p$$

Two-Group Case

In this case, sets of *N1* observations from group 1 and *N2* observations from group 2 are available on *p* response variables. We assume that all observations have the multivariate normal distribution with common variance covariance matrix $\varSigma$. The mean vectors of the two groups are assumed to be μ_1 and μ_2 under the alternative hypothesis. Under the null hypothesis, these mean vectors are assumed to be equal.

The value of *T*2 is computed using the formula

$$T^2_{p,N1+N2-2} = \frac{N1N2}{N1+N2}(\bar{y}_1 - \bar{y}_2)' S_{sp}^{-1}(\bar{y}_1 - \bar{y}_2)$$

where $\bar{y}_1$ and $\bar{y}_2$ are the vectors sample mean vectors of the two groups and S_{pl} is the pooled sample variance-covariance matrix.

To calculate power we need the non-centrality parameter for this distribution. This non-centrality parameter is defined as follows

$$\begin{aligned}\lambda &= \frac{N1N2}{N1+N2}(\mu_1 - \mu_2)' \varSigma^{-1}(\mu_1 - \mu_2) \\ &= \frac{N1N2}{N1+N2}\varDelta^2\end{aligned}$$

where

$$\Delta = \sqrt{(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)}$$

We define Δ as *effect size* because it provides a expression for the magnitude of the standardized difference between the null and alternative means.

Using this non-centrality parameter, the power of the Hotelling's T2 may be calculated for any value of the means and standard deviations. Since there is a simple relationship between the non-central T2 and the non-central *F*, calculations are actually based on the non-central *F* using the formula

$$\beta = \Pr\left(F' < F'_{\alpha,df1,df2,\lambda}\right)$$

where

$$df1 = p$$

$$df2 = N1 + N2 - p - 1$$

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data and Covariance tabs. To find out more about using the other tabs such as Axes/Legend/Grid, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains many of the options that you will be primarily concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for.

When you choose to solve for *N* (one-group) or *N1* (two-group), the program searches for the lowest sample size that meets the alpha and power criterion you have specified.

Error Rates

Power or Beta

This option specifies one or more values for power or for beta (depending on the chosen setting). Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of group one. For the one-group case, this is the value of *N*. For the two-group case, this is the value of *N*1.

You may enter a range of values such as '10 to 100 by 10'.

N2 (Sample Size Group 2)

In the two-group case, enter the value (or range of values) for the sample size of group two. In the one-group case, this value is ignored.

- **Use R**

  Enter 'Use R' if you want *N*2 to be calculated using the formula: *N*2=[*R* x *N*1] where *R* is the Sample Allocation Ratio and [*Y*] is the first integer >= *Y*. For example, if you want *N*1=*N*2, select 'Use R' here and set *R* equal to one.

R (Sample Allocation Ratio)

Enter a value (or range of values) for R, the allocation ratio between sample sizes. This value is only used when N2 is set to 'Use R'. When used, N2 is calculated from N1 using the formula: N2=[R x N1] where [Y] is the next integer >= Y. Note that setting R = 1.0 forces N2 = N1.

Note that this value is only used in the two-group case.

Groups

Number of Groups

Specify whether the analysis is for one group or two groups. If '1' is selected, N1 is used as the sample size (N) and the values of N2 and R are ignored.

Effect Size – Response Variables

Number of Response Variables

Enter the number of response (dependent or Y) variables. For a true multivariate test, this value will be greater than one.

The number of mean differences entered in the Mean Differences box or in the Means column must equal this value. If you read-in the covariance matrix from the spreadsheet, the number of columns specified must equal this value.

Effect Size – Mean Differences

Mean Differences (= # of Response Vars)

Enter a list of values representing the mean differences under the alternative hypothesis. Under the null hypothesis, these values are all zero. The values entered here represent the differences that you want the experiment (study) to be able to detect.

Note that the number of values must match the number of Response Variables.

If you like, you can enter these values in a column on the spreadsheet. This column is specified using the ‘Means Column’ option. When that option is specified, any values entered here are ignored.

Means Differences Column

Use this option to specify the spreadsheet column containing the hypothesized mean differences. The response variables are represented down the rows. The number of rows with data must equal the number of response variables. When this option is used, the 'Mean Differences' box is ignored.

You can obtain the spreadsheet by selecting ‘Window’, then ‘Data’, from the menus.

Effect Size – Mean Multiplier

K (Means Multipliers)

These values are multiplied times the mean differences to give you various effect sizes. A separate power calculation is generated for each value of K. If you want to ignore this setting, enter ‘1’.

Covariance Tab

This tab specifies the covariance matrix.

Covariance Matrix Specification

Specify Which Covariance Matrix Input Method to Use

This option specifies which method will be used to define the covariance matrix.

- **Standard Deviation and Correlation**

  This option generates a covariance matrix based on the settings for the standard deviation (SD) and the pattern of correlations as specified in the Correlation Pattern and R options.

- **Covariance Matrix Variables**

  When this option is selected, the covariance matrix is read in from the columns of the spreadsheet. This is the most flexible method, but specifying a covariance matrix is tedious. You will usually only use this method when a specific covariance is given to you.

Note that the spreadsheet is shown by selecting the menus: 'Window' and then 'Data'.

Covariance Matrix Specification-Input Method = 'Standard Deviation and Correlation'

The parameters in this section provide a flexible way to specify Σ , the covariance matrix. Because the covariance matrix is symmetric, it can be represented as

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_p\rho_{1p} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \cdots & \sigma_2\sigma_p\rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_p\rho_{1p} & \sigma_2\sigma_p\rho_{2p} & \cdots & \sigma_p^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{bmatrix}$$

where p is the number of response variables.

Thus, the covariance matrix can be represented with complete generality by specifying the standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_p$ and the correlation matrix

$$R = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}.$$

SD (Common Standard Deviation)

This value is used to generate the covariance matrix. This option specifies a single standard deviation to be used for all response variables. The square of this value becomes the diagonal elements of the covariance matrix. Since this is a standard deviation, it must be greater than zero.

This option is only used when the first Covariance Matrix Input Method is selected.

R (Correlation)

Specify a correlation to be used in calculating the off-diagonal elements of the covariance matrix. Since this is a correlation, it must be between -1 and 1. This option is only used when the first Covariance Matrix Input Method is selected.

Specify Correlation Pattern

This option specifies the pattern of the correlations in the variance-covariance matrix. Two options are available:

- **Constant**

  The value of R is used as the constant correlation. For example, if R = 0.6 and *p* = 6, the correlation matrix would appear as

$$R = \begin{bmatrix} 1 & 0.600 & 0.600 & 0.600 & 0.600 & 0.600 \\ 0.600 & 1 & 0.600 & 0.600 & 0.600 & 0.600 \\ 0.600 & 0.600 & 1 & 0.600 & 0.600 & 0.600 \\ 0.600 & 0.600 & 0.600 & 1 & 0.600 & 0.600 \\ 0.600 & 0.600 & 0.600 & 0.600 & 1 & 0.600 \\ 0.600 & 0.600 & 0.600 & 0.600 & 0.600 & 1 \end{bmatrix}$$

- **1st-Order Autocorrelation**

  The value of R is used as the base autocorrelation in a first-order, serial correlation pattern. For example, R = 0.6 and *p* = 6, the correlation matrix would appear as

$$R = \begin{bmatrix} 1 & 0.600 & 0.360 & 0.216 & 0.130 & 0.078 \\ 0.600 & 1 & 0.600 & 0.360 & 0.216 & 0.130 \\ 0.360 & 0.600 & 1 & 0.600 & 0.360 & 0.216 \\ 0.216 & 0.360 & 0.600 & 1 & 0.600 & 0.360 \\ 0.130 & 0.216 & 0.360 & 0.600 & 1 & 0.600 \\ 0.078 & 0.130 & 0.216 & 0.360 & 0.600 & 1 \end{bmatrix}$$

  This pattern is often chosen as the most realistic when little is known about the correlation pattern and the responses variables are measured across time.

Covariance Matrix Specification-Input Method = 'Covariance Matrix Variables'

This option instructs the program to read the covariance matrix from the spreadsheet.

Spreadsheet Columns Containing the Covariance Matrix

This option designates the columns on the current spreadsheet holding the covariance matrix. It is used when the 'Specify Which Covariance Matrix Input Method to Use' option is set to *Covariance Matrix Variables*. The number of columns and number of rows must match the number of response variable at which the subjects are measured.

Example 1 – Power in the One-Group Case and Validation

Rencher (1998) page 106 presents an example of power calculations for the one-group case in which the mean differences are both 1.88 and the covariance matrix is

$$\Sigma = \begin{bmatrix} 56.78 & 11.98 \\ 11.98 & 29.28 \end{bmatrix}$$

When N is 25 and the significance level is 0.05, Rencher calculated the power to be 0.3397.

To allow for a nice chart, we will calculate the power for several samples sizes and for K equal 1.0 and 1.5.

For your convenience, the covariance matrix has been stored in a spreadsheet called RENCHER2.S0. You must open that spreadsheet to run this example.

Setup

In order to run this example the **Rencher2.S0** data must be loaded into the spreadsheet. To open the spreadsheet window from the PASS Home Window, click on the **Tools** menu and select **Spreadsheet**. Once the spreadsheet is open, the **Rencher2.S0** data is loaded by clicking the **File** menu and selecting **Open**. The **Rencher2.S0** file is then selected from the **DATA** folder (the default location for this folder is *C:\...\[My] Documents\NCSS\PASS2008*). Then click **Open**.

This section presents the values of each of the parameters needed to run this example. From the PASS Home window, load the **Hotelling's T2** procedure window by clicking on **Means**, then **Multivariate Means**, then **Hotelling's T-Squared**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**5 15 25 35 50 75 100 150**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Number of Groups	**1**
Number of Response Variables	**2**
Mean Differences	**1.88 1.88**
Mean Differences Column	*blank*
K (Means Multiplier)	**1.0 1.5**
Covariance Tab	
Specify Covariance Method	**2) Covariance Matrix Variables**
Spreadsheet Columns	**VC1-VC2**

Reports Tab

Show Numeric Reports .... **Checked**
Show Means Matrix................................. **Checked**
Show Covariance Matrix .... **Checked**
Show Definitions .... **Checked**
Number of Summary Statements............ **1**
Show Plot .... **Checked**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Numeric Report

```
                    Multiply                                 Number
                       Means                        Effect    of Y's
Power          N          By     Alpha      Beta      Size     (DF1)      DF2
0.0737         5      1.0000    0.0500    0.9263      0.38         2        3
0.1996        15      1.0000    0.0500    0.8004      0.38         2       13
0.3379        25      1.0000    0.0500    0.6621      0.38         2       23
0.4707        35      1.0000    0.0500    0.5293      0.38         2       33
0.6409        50      1.0000    0.0500    0.3591      0.38         2       48
0.8311        75      1.0000    0.0500    0.1689      0.38         2       73
0.9282       100      1.0000    0.0500    0.0718      0.38         2       98
0.9895       150      1.0000    0.0500    0.0105      0.38         2      148
0.1040         5      1.5000    0.0500    0.8960      0.38         2        3
0.4033        15      1.5000    0.0500    0.5967      0.38         2       13
0.6635        25      1.5000    0.0500    0.3365      0.38         2       23
0.8302        35      1.5000    0.0500    0.1698      0.38         2       33
0.9475        50      1.5000    0.0500    0.0525      0.38         2       48
0.9943        75      1.5000    0.0500    0.0057      0.38         2       73
0.9995       100      1.5000    0.0500    0.0005      0.38         2       98
1.0000       150      1.5000    0.0500    0.0000      0.38         2      148

Report Definitions
Power is the probability of rejecting a false null hypothesis. Note that Power = 1 - Beta.
N is the sample size, the number of subjects in the experiment or study.
K is a constant by which all means are multiplied.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis. Note that Beta = 1 - Power.
Effect Size is a standardized version of T2 under the alternative hypothesis.
DF1 is the first degrees of freedom of T2. It is the number of response variables.
DF2 is the second degrees of freedom of T2.

Summary Statements
A sample size of 5 achieves 7% power to detect an effect size of 0.38 which represents the
differences between the null and alternative means of the 2 response variables, adjusted by the
variance-covariance matrix. The one-sample Hotelling's T-squared test statistic is used with a
significance level of 0.0500.
```

This report gives the power for each value of *N* and *K*. Notice that the power for *K* = 1 and *N* = 25 is 0.3379. This is slightly different than the 0.3397 obtained by interpolation by Rencher.

Means Matrix

```
Means Matrix Section
Name             Mean
Y1             1.8800
Y2             1.8800
```

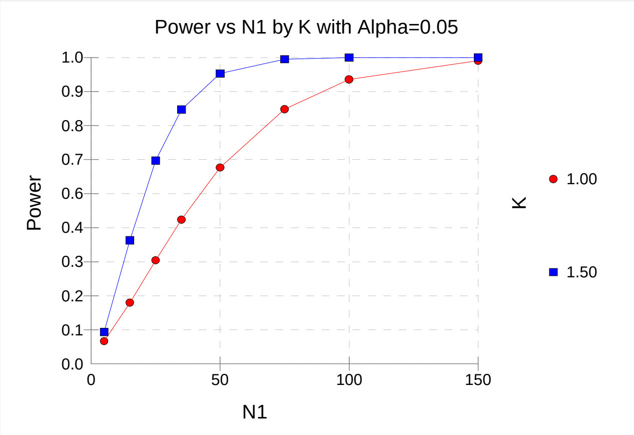
This report shows the mean differences that were read in. When a Means Multiplier, *K*, is used, each value of *K* is multiplied times each of these values.

Variance-Covariance Matrix Section

```
Variance-Covariance Matrix Section

Response           Y1           Y2
Y1             7.5353       0.2938
Y2             0.2938       5.4111
```

This report shows the variance-covariance matrix that was read in from the spreadsheet or generated by the settings of on the Covariance tab. The standard deviations are given on the diagonal and the correlations are given off the diagonal.

Chart Section



This chart shows the relationship between power and *N* for each value of *K*.

Example 2 – Power in the Two-Group Case and Validation

Rencher (1998) pages 107-108 presents an example of power calculations for the two-group case in which the mean differences and covariance matrix are

$$\mu_1 - \mu_2 = \begin{bmatrix} 3 \\ -2 \\ 3 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 6 & -3 & 3 \\ -3 & 5 & -6 \\ 3 & -6 & 9 \end{bmatrix}$$

When $N1 = N2 =$ 10, 12, 14, 16 and the significance level is 0.05, Rencher calculated the power to be 0.6438, 0.7520, 0.8329, 0.8936, respectively.

For your convenience, the mean differences and covariance matrix have been stored in a spreadsheet called RENCHER2.S0. You must open that spreadsheet to run this example.

Setup

In order to run this example the **Rencher2.S0** data must be loaded into the spreadsheet. To open the spreadsheet window from the PASS Home Window, click on the **Tools** menu and select **Spreadsheet**. Once the spreadsheet is open, the **Rencher2.S0** data is loaded by clicking the **File** menu and selecting **Open**. The **Rencher2.S0** file is then selected from the **DATA** folder (the default location for this folder is *C:\...\[My] Documents\NCSS\PASS2008*). Then click **Open**.

This section presents the values of each of the parameters needed to run this example. From the PASS Home window, load the **Hotelling's T2** procedure window by clicking on **Means**, then **Multivariate Means**, then **Hotelling's T-Squared**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
Alpha	**0.05**
N1 (Sample Size Group 1)	**10 12 14 16**
N2 (Sample Size Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
Number of Groups	**2**
Number of Response Variables	**3**
Mean Differences	*blank*
Mean Differences Column	**Differences**
K (Means Multiplier)	**1.0**
Covariance Tab	
Specify Covariance Method	**2) Covariance Matrix Variables**
Spreadsheet Columns	**VC_1-VC_3**

Reports Tab

Show Numeric Results............................**Checked**
Show Means Matrix.................................**Checked**
Show Covariance Matrix ....**Checked**
Show Definitions ....**Checked**
Number of Summary Statements............**1**
Show Plot ....**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Report

		Multiply				Number		
		Means			Effect	of Y's		
Power	N	By	Alpha	Beta	Size	(DF1)	DF2	
0.6442	10	10	1.0000	0.0500	0.3558	1.41	3	16
0.7546	12	12	1.0000	0.0500	0.2454	1.41	3	20
0.8361	14	14	1.0000	0.0500	0.1639	1.41	3	24
0.8936	16	16	1.0000	0.0500	0.1064	1.41	3	28

Note that the power values obtained here are very close to those obtained by Rencher. We feel that our results are more accurate since Rencher’s results were obtained by interpolation from Tang’s tables.

Chapter 605

Multivariate Analysis of Variance (MANOVA)

Introduction

This module calculates power for multivariate analysis of variance (MANOVA) designs having up to three factors. It computes power for three MANOVA test statistics: Wilks' lambda, Pillai-Bartlett trace, and Hotelling-Lawley trace.

MANOVA is an extension of common analysis of variance (ANOVA). In ANOVA, differences among various group means on a single-response variable are studied. In MANOVA, the number of response variables is increased to two or more. The hypothesis concerns a comparison of vectors of group means. The multivariate extension of the *F*-test is not completely direct. Instead, several test statistics are available. The actual distributions of these statistics are difficult to calculate, so we rely on approximations based on the *F*-distribution.

Assumptions

The following assumptions are made when using MANOVA to analyze a factorial experimental design.

1. The response variables are continuous.
2. The residuals follow the multivariate normal probability distribution with mean zero and constant variance-covariance matrix.
3. The subjects are independent.

Technical Details

General Linear Multivariate Model

This section provides the technical details of the MANOVA designs that can be analyzed by ***PASS***. The approximate power calculations outlined in Muller, LaVange, Ramey, and Ramey (1992) are used. Using their notation, for *N* subjects, the usual general linear multivariate model is

$$\underset{(N\times p)}{Y} = \underset{(N\times q\times p)}{XM} + \underset{(N\times p)}{R}$$

where each row of the residual matrix *R* is distributed as a multivariate normal

$$row_k(R) \sim N_p(0,\Sigma)$$

Note that *p* is the number of response variables and *q* is the number of design variables, *Y* is the matrix of responses, *X* is the design matrix, *M* is the matrix of regression parameters (means), and *R* is the matrix of residuals.

Hypotheses about various sets of regression parameters are tested using

$$H_0\colon \underset{a\times p}{\Theta} = \Theta_0$$

$$\underset{a\times q\times p}{CM} = \Theta$$

where *C* is an orthonormal contrast matrix and Θ_0 is a matrix of hypothesized values, usually zeros. Note that *C* defines contrasts among the factor levels. Tests of the various main effects and interactions may be constructed with suitable choices of *C*. These tests are based on

$$\hat{M} = \left(X'X\right)^{-}X'Y$$

$$\hat{\Theta} = C\hat{M}$$

$$\underset{p\times p}{H} = \left(\hat{\Theta}-\Theta_0\right)'\left[C\left(X'X\right)^{-}C'\right]^{-1}\left(\hat{\Theta}-\Theta_0\right)$$

$$\underset{p\times p}{E} = \hat{\Sigma}\cdot\left(N-r\right)$$

$$\underset{p\times p}{T} = H+E$$

where *r* is the rank of *X*.

Wilks' Lambda Approximate F Test

The hypothesis $H_0\colon\Theta=\Theta_0$ may be tested using Wilks' likelihood ratio statistic *W*. This statistic is computed using

$$W = \left|ET^{-1}\right|$$

An F approximation to the distribution of W is given by

$$F_{df_1,df_2} = \frac{\eta / df_1}{(1-\eta) / df_2}$$

where

$$\lambda = df_1 F_{df_1,df_2}$$

$$\eta = 1 - W^{1/g}$$

$$df1 = ap$$

$$df2 = g\left[(N-r) - (p-a+1)/2\right] - (ap-2)/2$$

$$g = \left(\frac{a^2 p^2 - 4}{a^2 + p^2 - 5}\right)^{\frac{1}{2}}$$

Pillai-Bartlett Trace Approximate F Test

The hypothesis $H_0\!:\Theta = \Theta_0$ may be tested using the Pillai-Bartlett Trace. This statistic is computed using

$$T_{PB} = tr\left(HT^{-1}\right)$$

A noncentral F approximation to the distribution of T_{PB} is given by

$$F_{df_1,df_2} = \frac{\eta / df_1}{(1-\eta) / df_2}$$

where

$$\lambda = df_1 F_{df_1,df_2}$$

$$\eta = \frac{T_{PB}}{s}$$

$$s = \min(a, p)$$

$$df1 = ap$$

$$df2 = s\left[(N-r) - p + s\right]$$

Hotelling-Lawley Trace Approximate F Test

The hypothesis $H_0\!:\Theta = \Theta_0$ may be tested using the Hotelling-Lawley Trace. This statistic is computed using

$$T_{HL} = tr\left(HE^{-1}\right)$$

An *F* approximation to the distribution of T_{HL} is given by

$$F_{df_1,df_2} = \frac{\eta / df_1}{(1-\eta)/df_2}$$

where

$$\lambda = df_1 F_{df_1,df_2}$$

$$\eta = \frac{\frac{T_{HL}}{s}}{1+\frac{T_{HL}}{s}}$$

$$s = \min(a, p)$$

$$df\,1 = ap$$

$$df\,2 = s\left[(N-r)-p+s\right]$$

M (Mean) Matrix

In the general linear multivariate model presented above, *M* represents a matrix of regression coefficients. Although other structures and interpretations of *M* are possible, in this module we assume that the elements of *M* are the cell means. The rows of *M* represent the factor categories and the columns of *M* represent the response variables. (Note that this is just the opposite of the orientation used when entering *M* into the spreadsheet.)

The *q* rows of *M* represent the *q* groups into which the subjects can be classified. For example, if a design includes three factors with 2, 3, and 4 categories, the matrix *M* would have 2 *x* 3 *x* 4 = 24 rows. That is, *q* = 24.

Consider now an example in which *q* = 3 and *p* = 4. That is, there are three groups into which subjects can be placed. Each subject has four made. The matrix *M* would appear as follows.

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{bmatrix}$$

For example, the element μ_{12} is the mean of the second response of subjects in the first group. To calculate the power of this design, you would need to specify appropriate values of all twelve means.

C Matrix – Contrasts

The *C* matrix is comprised of contrasts that are applied to the rows of *M*. You do not have to specify these contrasts. They are generated for you. You should understand that a different *C* matrix is generated for each term in the model.

Generating the C Matrix when there are Multiple Between Factors

Generating the *C* matrix when there is more than one factor is more difficult. We use the method of O'Brien and Kaiser (1985) which we briefly summarize here.

Step 1. Write a complete set of contrasts suitable for testing each factor separately. For example, if you have three factors with 2, 3, and 4 categories, you might use

$$\ddot{C}_{B1} = \begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},\ \ddot{C}_{B2} = \begin{bmatrix} \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},\ \text{and}\ \ddot{C}_{B3} = \begin{bmatrix} \frac{-3}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} \\ 0 & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Step 2. Define appropriate J_k matrices corresponding to each factor. These matrices comprised of one row and *k* columns whose equal element is chosen so that the sum of its elements squared is one. In this example, we use

$$J_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},\ J_3 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix},\ J_4 = \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix}$$

Step 3. Create the appropriate contrast matrix using a direct (Kronecker) product of either the $\ddot{C}_{Bi}$ matrix if the factor is included in the term or the J_i matrix when the factor is not in the term. Remember that the direct product is formed by multiplying each element of the second matrix by all members of the first matrix. Here is an example

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 & 0 & 0 & -1 & -2 \\ 3 & 4 & 0 & 0 & -3 & -4 \\ 0 & 0 & 2 & 4 & 0 & 0 \\ 0 & 0 & 6 & 8 & 0 & 0 \\ -1 & -2 & 0 & 0 & 3 & 6 \\ -3 & -4 & 0 & 0 & 9 & 12 \end{bmatrix}$$

As an example, we will compute the *C* matrix suitable for testing factor *B*2

$$C_{B2} = J_2 \otimes \ddot{C}_{B2} \otimes J_4$$

Expanding the direct product results in

$$
\begin{aligned}
C_{B2} &= J_2 \otimes \ddot{C}_{B2} \otimes J_4 \\
&= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{-2}{\sqrt{12}} & \frac{-2}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} \\ 0 & 0 & \frac{-1}{\sqrt{4}} & \frac{-1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix} \otimes \begin{bmatrix} \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} & \frac{1}{\sqrt{4}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{-2}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} & \frac{1}{\sqrt{48}} \\ 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & 0 & 0 & \frac{-1}{\sqrt{16}} & \frac{-1}{\sqrt{16}} & \frac{1}{\sqrt{16}} & \frac{1}{\sqrt{16}} \end{bmatrix}
\end{aligned}
$$

Similarly, the C matrix suitable for testing interaction $B2B3$ is

$$C_{B2B3} = J_2 \otimes \ddot{C}_{B2} \otimes \ddot{C}_{B3}$$

We leave the expansion of this matrix *PASS*, but we think you have the idea.

Power Calculations

To calculate statistical power, we must determine distribution of the test statistic under the alternative hypothesis which specifies a different value for the regression parameter matrix B. The distribution theory in this case has not been worked out, so approximations must be used. We use the approximations given by Muller and Barton (1989) and Muller, LaVange, Ramey, and Ramey (1992). These approximations state that under the alternative hypothesis, F_U is distributed as a noncentral F random variable with degrees of freedom and noncentrality shown above. The calculation of the power of a particular test may be summarized as follows.

1. Specify values of X, M, Σ, C, and Θ_0.
2. Determine the critical value using $F_{crit} = FINV\left(1-\alpha, df1, df2\right)$, where *FINV*( ) is the inverse of the central F distribution and α is the significance level.
3. Compute the noncentrality parameter λ.
4. Compute the power as

   $$Power = 1 - NCFPROB\left(F_{crit}, df1, df2, \lambda\right)$$

   where *NCFPROB*( ) is the noncentral F distribution.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data and Covariance tabs. To find out more about using the other tabs such as Axes, Plot Text, or Template, go to the Procedure Window chapter.

Data Tab

The Data tab contains many of the options that you will be primarily concerned with.

Solve For

Find (Solve For)

This option specifies the parameter to be solved for. If you choose to solve for *n* (sample size), you must also specify which test statistic you want to use.

When you choose to solve for *n*, the program searches for the lowest sample size that meets the alpha and beta criterion you have specified for each of the terms. If you do not want a term to be used in the search, set its alpha and beta values to 0.99.

Also, when the '= n's' box is not checked, the search is made using unequal group sample sizes. The relative proportion of the sample in each group is set by the values of *n* given in the Subjects Per Group box. For example, if your design has three groups and you entered '1 1 2' in the Subjects Per Group box, the search will only consider designs in which the size of the last group is twice the rest. That is, it will consider '2 2 4', '3 3 6', '4 4 8', etc.

Note: no plots are generated when you solve for *n*.

Sample Size

n (Subjects Per Group)

Specify one or more values for the number of subjects per group. The total sample size is the sum of the individual group sizes across all groups.

You can specify a list like '2 4 6'. The items in the list may be separated with commas or blanks. The interpretation of the list depends on the =n's check box. When the =n's box is checked, a separate analysis is calculated for each value of *n*. When the =n's box is not checked, ***PASS*** uses the *n's* as the actual group sizes. In this case, the number of items entered must match the number of groups in the design.

When you choose to solve for *n* and the = n's box is not checked, the search is made using unequal group sample sizes. The relative proportion of the sample in each group is set by the values of *n* given in this box. For example, if your design has three groups and you enter '1 1 2' here, the search will only consider designs in which the size of the last group is twice the rest. That is, it will consider '2 2 4', '3 3 6', '4 4 8', etc.

= n's

This option controls whether the number of subjects per group is to be equal for all groups or not. When checked, the number of subjects per group is equal for all groups. A list of values such as '5 10 15' represents three designs: one with five per group, one with ten per group, and one with fifteen per group.

When this option is not checked, the *n*'s are assumed to be unequal. A list of values represents the size of the individual groups. For example, '5 10 15' represents a single, three-group design with five in the first group, ten in the second group, and fifteen in the third group.

Effect Size – Response Variables

Number of Response Variables

Enter the number of response variables in your design. For a true MANOVA, this value must be greater than one. The number of rows in the means matrix must equal this value. If you specify a covariance matrix, the number of columns specified must equal this value.

Effect Size – Means

Means Matrix

Use this option to the specify spreadsheet columns containing a hypothesized means matrix that is used to specify the alternative hypothesis. You can obtain the spreadsheet by selecting 'Window', then 'Data', from the menus.

The factors are represented across the columns of the spreadsheet and the response variables are represented down the rows. The number of columns specified must equal the number of groups. The number of rows with data in these columns must equal the number of response variables. For example, suppose you are designing an experiment that is to have two factors (A and B) and two response variables (Y1 and Y2). Suppose each of the factors has two levels. The four columns of the spreadsheet would represent

```
A1B1 A1B2 A2B1 A2B2.
```

The two rows of the spreadsheet would represent

```
Y1
Y2
```

K (Means Multipliers)

These values are multiplied times the means matrix to give you various effect sizes. A separate power calculation is generated for each value of K. These values become the horizontal axis in the second power chart. If you want to ignore this setting, enter '1'.

Effect Size – Main Effects & Interactions

Labels

Specify a label for this factor. Although we suggest that only a single letter be used, the label can consist of several letters. When several letters are used, the labels for the interactions may be extra long and confusing. Of course, you must be careful not to use the same label for two factors.

One of the easiest sets of labels is to use A, B, and C for the factors.

Levels

Specify the number of levels (categories) in this factor. Typical values are from 2 to 8. Set this to a blank (or 0) to ignore the factor in the design.

Alpha

These options specify the probability of a type-I error (alpha) for each factor and interaction. A type-I error occurs when you reject the null hypothesis of zero effects when in fact they are zero. Since they are probabilities, alpha values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This value may be interpreted as meaning that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You can specify different alpha values for different terms. For example, although you have three terms in an experiment, you might be mainly interested in only one of them. Hence, you could increase the alpha level of the tests of the other terms and thereby increase their power. Also, you may want to increase the alpha level of the interaction terms, as these will often have poor power otherwise.

Power or Beta

These options specify the power or for beta (depending on the chosen setting) for each factor and interaction.

Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Covariance Tab

This tab specifies the covariance matrix.

Covariance Matrix Specification

Specify Which Covariance Matrix Input Method to Use

This option specifies which method will be used to define the covariance matrix.

- **Standard Deviation and Correlation**

  This option generates a covariance matrix based on the settings for the standard deviation (SD) and the pattern of correlations as specified in the Correlation Pattern and R options.

- **Covariance Matrix Variables**

  When this option is selected, the covariance matrix is read in from the columns of the spreadsheet. This is the most flexible method, but specifying a covariance matrix is tedious. You will usually only use this method when a specific covariance is given to you.

  Note that the spreadsheet is shown by selecting the menus: ‘Window’ and then ‘Data’.

Covariance Matrix Specification- Input Method = 'Standard Deviation and Correlation'

The parameters in this section provide a flexible way to specify Σ , the covariance matrix. Because the covariance matrix is symmetric, it can be represented as

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_p\rho_{1p} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \cdots & \sigma_2\sigma_p\rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_p\rho_{1p} & \sigma_2\sigma_p\rho_{2p} & \cdots & \sigma_p^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{bmatrix}$$

where p is the number of response variables.

Thus, the covariance matrix can be represented with complete generality by specifying the standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_p$ and the correlation matrix

$$R = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}.$$

SD (Standard Deviation)

This value is used to generate the covariance matrix. This option specifies a standard deviation to be used for all response variables. The square of this value becomes the diagonal elements of the covariance matrix. Since this is a standard deviation, it must be greater than zero.

This option is only used when the first Covariance Matrix Input Method is selected.

R (Correlation)

Specify a correlation to be used in calculating the off-diagonal elements of the covariance matrix. Since this is a correlation, it must be between -1 and 1. This option is only used when the first Covariance Matrix Input Method is selected.

Specify Correlation Pattern

This option specifies the pattern of the correlations in the variance-covariance matrix. Two options are available:

- **Constant**

  The value of R is used as the constant correlation. For example, if R = 0.6 and *p* = 6, the correlation matrix would appear as

$$R = \begin{bmatrix} 1 & 0.600 & 0.600 & 0.600 & 0.600 & 0.600 \\ 0.600 & 1 & 0.600 & 0.600 & 0.600 & 0.600 \\ 0.600 & 0.600 & 1 & 0.600 & 0.600 & 0.600 \\ 0.600 & 0.600 & 0.600 & 1 & 0.600 & 0.600 \\ 0.600 & 0.600 & 0.600 & 0.600 & 1 & 0.600 \\ 0.600 & 0.600 & 0.600 & 0.600 & 0.600 & 1 \end{bmatrix}$$

- **1st Order Autocorrelation**

  The value of R is used as the base autocorrelation in a first-order, serial correlation pattern. For example, R = 0.6 and *p* = 6, the correlation matrix would appear as

$$R = \begin{bmatrix} 1 & 0.600 & 0.360 & 0.216 & 0.130 & 0.078 \\ 0.600 & 1 & 0.600 & 0.360 & 0.216 & 0.130 \\ 0.360 & 0.600 & 1 & 0.600 & 0.360 & 0.216 \\ 0.216 & 0.360 & 0.600 & 1 & 0.600 & 0.360 \\ 0.130 & 0.216 & 0.360 & 0.600 & 1 & 0.600 \\ 0.078 & 0.130 & 0.216 & 0.360 & 0.600 & 1 \end{bmatrix}$$

  This pattern is often chosen as the most realistic when little is known about the correlation pattern and the responses variables are measured across time.

Covariance Matrix Specification-Input Method = ‘Covariance Matrix Variables’

This option instructs the program to read the covariance matrix from the spreadsheet.

Spreadsheet Columns Containing the Covariance Matrix

This option designates the columns on the current spreadsheet holding the covariance matrix. It is used when the ‘Specify Which Covariance Matrix Input Method to Use’ option is set to *Covariance Matrix Variables*. The number of columns and number of rows must match the number of response variable at which the subjects are measured.

Reports Tab

This tab specifies which reports and graphs are displayed as well as their format.

Select Output – Numeric Reports

Test in Summary Statement(s)

Indicate the test that is to be reported on in the Summary Statements.

Report Options

Maximum Term-Order Reported

Indicate the maximum order of terms to be reported on. Occasionally, higher-order interactions are of little interest and so they may be omitted. For example, enter a ‘2’ to limit output to individual factors and two-way interactions.

Skip Line After

The names of the terms can be too long to fit in the space provided. If the name contains more characters than this, the rest of the output is placed on a separate line. Enter ‘1’ when you want every term’s results printed on two lines. Enter ‘100’ when you want every variable’s results printed on one line.

Example 1 – Determining Power

Researchers are planning a study of the impact of a drug. They want to evaluate the differences in heart rate and blood pressure among three age groups: 20-40, 41-60, and over 60. They want to be able to detect a 10% change in heart rate and in blood pressure among the age groups. They plan to analyze the data using Wilks’ lambda.

Previous studies have found an average heart rate of 93 with a standard deviation of 4 and an average blood pressure of 130 with a standard deviation of 5. The correlation between the two responses will be set at 0.7.

From a heart rate of 93, a 10% reduction gives 84. They want to be able to detect age-group heart-rate means the range from 93 to 84. From a blood pressure of 130, a 10% reduction gives 117. The want to be able to detect age-group blood-pressure means that range from 130 to 117. Hence, the means matrix that they will use is

C1	C2	C3
93	88	84
130	124	117

Base on the standard deviation settings that they chose to use, the covariance matrix will be

C4	C5
16	14
14	25

In order to understand the relationship between power and sample size, they decide to calculate power values for sample sizes between 2 and 12, using a 0.05 significance level.

For your convenience, the Means Matrix and Covariance Matrix have been stored in a spreadsheet called PASSMANOVA1.S0. You can enter the above values yourself or you can open that spreadsheet.

Setup

In order to run this example the **PASSMANOVA1.S0** data must be loaded into the spreadsheet. To open the spreadsheet window from the PASS Home Window, click on the **Tools** menu and select **Spreadsheet**. Once the spreadsheet is open, the **PASSMANOVA1.S0** data is loaded by clicking the **File** menu and selecting **Open**. The **PASSMANOVA1.S0** file is then selected from the **DATA** folder (the default location for this folder is *C:\...\[My] Documents\NCSS\PASS2008*). Then click **Open**.

This section presents the values of each of the parameters needed to run this example. From the PASS Home window, load the **Multivariate Analysis of Variance (MANOVA)** procedure window by clicking on **Means**, then **Multivariate Means**, then **Multivariate Analysis of Variance (MANOVA)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**2 4 6 8 10 12**
=n's	**checked**
Number of Response Variables	**2**
Means Matrix	**C1-C3**
K (Means Multipliers)	**0.5 1 1.5**
For Factor F1	
Label	**A**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Covariance Tab	
Specify Covariance Method	**2) Covariance Matrix Variables**
Spreadsheet Columns	**C4-C5**
Reports Tab	
Numeric Results by Term	**Checked**
Numeric Results by Design	**Not Checked**
Wilks' Lambda	**Checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
Means Matrix	**Checked**
Covariance Matrix	**Checked**
Test in Summary Statement	**Wilks Lambda**
Show Plot 1	**Checked**
Show Plot 2	**Checked**
Test That is Plotted	**Wilks Lambda**
Max Term-Order Plotted	**2**
Max Term-Order Reported	**2**

Annotated Output

Click the Run button to perform the calculations and generate the following output.

Term Report

```
Results for Factor A (Levels = 3)
                                 Multiply                Approx.
                                    Means       Test           F
Test       Power    n       N          By  Statistic   Statistic  DF1|DF2     Alpha       Beta
Wilks     0.0729    2       6        0.50     0.2115        0.27      4|4    0.0500     0.9271
Wilks     0.1291    2       6        1.00     0.4654        0.87      4|4    0.0500     0.8709
Wilks     0.2046    2       6        1.50     0.6185        1.62      4|4    0.0500     0.7954
Wilks     0.1888    4      12        0.50     0.1562        0.74     4|16    0.0500     0.8112
Wilks     0.5749    4      12        1.00     0.3853        2.51     4|16    0.0500     0.4251
Wilks     0.8722    4      12        1.50     0.5443        4.78     4|16    0.0500     0.1278
Wilks     0.3191    6      18        0.50     0.1437        1.17     4|28    0.0500     0.6809
Wilks     0.8548    6      18        1.00     0.3648        4.02     4|28    0.0500     0.1452
Wilks     0.9916    6      18        1.50     0.5241        7.71     4|28    0.0500     0.0084
Wilks     0.4488    8      24        0.50     0.1382        1.60     4|40    0.0500     0.5512
Wilks     0.9603    8      24        1.00     0.3554        5.51     4|40    0.0500     0.0397
Wilks     0.9997    8      24        1.50     0.5147       10.61     4|40    0.0500     0.0003
Wilks     0.5678   10      30        0.50     0.1351        2.03     4|52    0.0500     0.4322
Wilks     0.9907   10      30        1.00     0.3500        7.00     4|52    0.0500     0.0093
Wilks     1.0000   10      30        1.50     0.5092       13.49     4|52    0.0500     0.0000
Wilks     0.6704   12      36        0.50     0.1331        2.46     4|64    0.0500     0.3296
Wilks     0.9981   12      36        1.00     0.3465        8.48     4|64    0.0500     0.0019
Wilks     1.0000   12      36        1.50     0.5057       16.37     4|64    0.0500     0.0000

Summary Statements
A MANOVA design with 1 factor and 2 response variables has 3 groups with 2 subjects each for a
total of 6 subjects. This design achieves 7% power to test factor A if a Wilks' Lambda
Approximate F Test is used with a 5% significance level.
```

This report gives the power for each value of *n* and *K*. It is useful when you want to compare the powers of the terms in the design at a specific sample size.

In this example, for $K = 1$, the design goal of 0.95 power is achieved for $n = 8$.

The definitions of each of the columns of the report are as follows.

Test

This column identifies the test statistic. Since the power depends on the test statistic, you should make sure that this is the test statistic that you will use in your analysis.

Power

This is the computed power for the term.

n

The value of *n* is the number of subjects per group.

N

The value of *N* is the total number of subjects in the study.

Multiply Means By

This is the value of the means multiplier, *K*.

Test Statistic

This is the value of the test statistic computed at the hypothesized values. The name of the statistic is identified in the Test column. Possible values are Wilks' lambda, Pillai-Bartlett trace, or Hotelling-Lawley trace. The actual formulas used were given earlier in the Technical Details section.

Approx. F Statistic

This is the value of the *F* statistic that is used to compute the probability levels. This value is calculated using the hypothesized values. The actual formulas used were given earlier in the Technical Details section.

DF1|DF2

These are the numerator and denominator degrees of freedom of the approximating *F* distribution.

Alpha

Alpha is the significance level of the test.

Beta

Beta is the probability of failing to reject the null hypothesis when the alternative hypothesis is true.

Means Matrix

```
Means Matrix Section
Name              A1          A2          A3
Y1             93.00       88.00       84.00
Y2            130.00      124.00      117.00
```

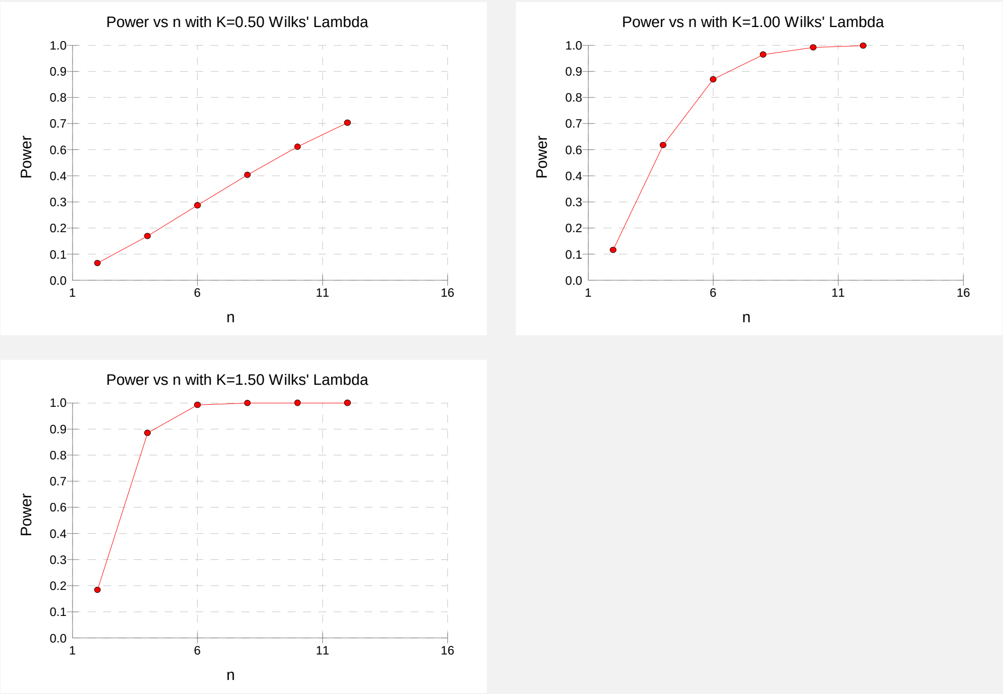
This report shows the means matrix that was read in. It may be used to get an impression of the magnitude of the difference among the means that is being studied. When a Means Multiplier, *K*, is used, each value of *K* is multiplied times each value of this matrix.

Variance-Covariance Matrix Section

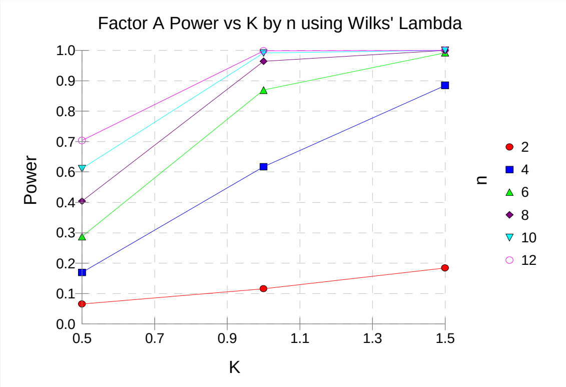
```
Variance-Covariance Matrix Section

Response          Y1          Y2
Y1              4.00        0.70
Y2              0.70        5.00
```

This report shows the variance-covariance matrix that was read in from the spreadsheet or generated by the settings of on the Covariance tab. The standard deviations are given on the diagonal and the correlations are given off the diagonal.

Plots Section



These chart show the relationship between power and *n* for each value of *K*. Note that high-order interactions may be omitted from the plot by reducing the Maximum Term-Order Plotted option on the Reports tab.



These charts show the relationship between power and *K* for each value of *n*. Remember that *K* is the mean multiplier. It changes the effect size.

Example 2 – Validation

In this example, we will set p = 2, q = 3, alpha = 0.05, and n = 4. The mean and covariance matrices are

$$M = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$$

The contrast matrix C is

$$C = \begin{bmatrix} \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

The $X'X$ matrix is

$$X'X = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

The matrix Θ is

$$\begin{aligned} \Theta &= CM \\ &= \begin{bmatrix} \frac{3}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \end{aligned}$$

The matrix H is

$$\begin{aligned} H &= \left(\hat{\Theta} - \Theta_0\right)' \left[C\left(X'X\right)^{-} C'\right]^{-1} \left(\hat{\Theta} - \Theta_0\right) \\ &= \begin{bmatrix} 8 & 4 \\ 4 & 8/3 \end{bmatrix} \end{aligned}$$

The matrix E is

$$\begin{aligned} E &= \hat{\Sigma} \cdot \left(N - r\right) \\ &= \begin{bmatrix} 36 & 9 \\ 9 & 36 \end{bmatrix} \end{aligned}$$

The matrix T is

$$\begin{aligned} T &= H + E \\ &= \begin{bmatrix} 44 & 13 \\ 13 & 38\tfrac{2}{3} \end{bmatrix} \end{aligned}$$

Using these matrices, we can calculate the values of the test statistics. We will only calculate the results for Wilks' lambda. We have

$$
\begin{aligned}
W &= \det\left(ET^{-1}\right) \\
&= 0.79290842 \\
a &= q-1 \\
&= 2 \\
g &= \left(\frac{a^2p^2-4}{a^2+p^2-5}\right)^{\frac{1}{2}} \\
&= \left(\frac{2^22^2-4}{2^2+2^2-5}\right)^{\frac{1}{2}} \\
&= 2 \\
\eta &= 1-W^{1/g} \\
&= 1-\sqrt{0.79290842} \\
&= 0.10954595 \\
df1 &= ap \\
&= 4 \\
df2 &= g\left[\left(N-r\right)-\left(p-a+1\right)/2\right]-\left(ap-2\right)/2 \\
&= 2\left[\left(12-3\right)-\left(2-2+1\right)/2\right]-\left(4-2\right)/2 \\
&= 16 \\
F_{df_1,df_2} &= \frac{\eta/df_1}{\left(1-\eta\right)/df_2} \\
&= \frac{0.10954595/4}{\left(1-0.10954595\right)/16} \\
&= 0.49209030 \\
\lambda &= df_1F_{df_1,df_2} \\
&= 4\left(0.49209030\right) \\
&= 1.96836120
\end{aligned}
$$

For an F with 4 and 16 degrees of freedom, the 5% critical value is 3.0069172799. Finally, compute the power using a noncentral F with 4 and 16 degrees of freedom and noncentrality parameter

$$
\begin{aligned}
Power &= \Pr\left(f > F|df_1 = 4, df_2 = 16, \lambda = 1.96836120\right) \\
&= 0.1370631884
\end{aligned}
$$

In order to run this example in ***PASS***, the values of the means and the covariance matrix (given above) must be entered on a spreadsheet. We have loaded these values into the database called PASSMANOVA2. Either enter the values yourself, or load the PASSMANOVA2 database which is located in the Data directory. The instructions below assume that the means are in columns C1-C3, while the covariance matrix is in columns C4-C5.

Setup

In order to run this example the **PASSMANOVA2.S0** data must be loaded into the spreadsheet. To open the spreadsheet window from the PASS Home Window, click on the **Tools** menu and select **Spreadsheet**. Once the spreadsheet is open, the **PASSMANOVA2.S0** data is loaded by clicking the **File** menu and selecting **Open**. The **PASSMANOVA2.S0** file is then selected from the **DATA** folder (the default location for this folder is *C:\...\[My] Documents\NCSS\PASS2008*). Then click **Open**.

This section presents the values of each of the parameters needed to run this example. From the PASS Home window, load the **Multivariate Analysis of Variance (MANOVA)** procedure window by clicking on **Means**, then **Multivariate Means**, then **Multivariate Analysis of Variance (MANOVA)**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
n (Subjects Per Group)	**4**
=n’s	**checked**
Number of Response Variables	**2**
Means Matrix	**C1-C3**
K (Means Multipliers)	**1**
For Factor F1	
Label	**A**
Levels	**3**
Alpha	**0.05**
Power	*Ignored since this is the Find setting*
Covariance Tab	
Specify Covariance Method	**2) Covariance Matrix Variables**
Spreadsheet Columns	**C4-C5**
Reports Tab	
Numeric Results by Term	**Checked**
Numeric Results by Design	**Not Checked**
Wilks’ Lambda	**Checked**
Pillai-Bartlett	**Not checked**
Hotelling-Lawley	**Not checked**
Means Matrix	**Checked**
Covariance Matrix	**Checked**

Output

Click the Run button to perform the calculations and generate the following output.

Term Report

```
Results for Factor A (Levels = 3)
                                    Multiply                  Approx.
                                       Means        Test            F
Test        Power      n       N          By   Statistic    Statistic  DF1|DF2      Alpha       Beta
Wilks      0.1371      4      12      1.0000      0.7929       0.4921     4|16     0.0500     0.8629
```

As you can see, the power computed here matches the results we computed manually.

Chapter 610

One-Sample or Paired T-Test for Microarray Data

Introduction

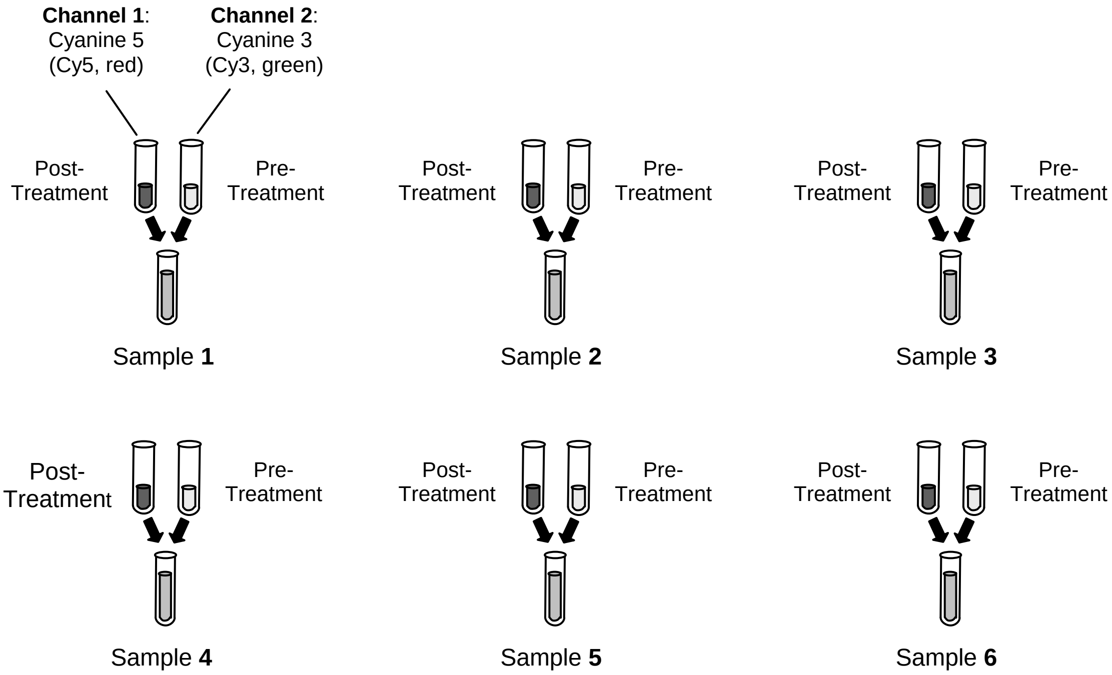
This chapter describes how to estimate power and sample size (number of arrays) for paired and one sample microarray studies using the PASS: One-Sample or Paired T-Test for Microarray Data procedure. False discovery rate and experiment-wise error rate control methods are available in this procedure. Values that can be varied in this procedure are power, false discovery rate and experiment-wise error rate, sample size (number of arrays), the minimum mean difference detected, the standard deviation, and in the case of false discovery rate control, the number of genes with minimum mean difference.

Paired Design (Two-Channel Arrays)

The paired design is often used in two-channel experiments when the gene expression comparison to be made involves a natural pairing of experimental units.

As an example, suppose 6 cell samples will be available for comparison. A portion of each of the 6 cell samples (before treatment) is to be reserved as a control. The same treatment will then be given to each of the 6 remaining portions of the samples. It is of interest to determine the genes that are differentially expressed when the treatment is given. In this scenario there is a natural before/after treatment pairing for each sample. The reserved control portions of each sample will be labeled with Cyanine 3 (Cy3, green) dye, while the treatment portions are to be labeled with Cyanine 5 (Cy5, red) dye. From each sample, the labeled control and the labeled treatment portions will be mixed and exposed to an array. The control and treatment portions compete to bind at each spot. The expression of treatment and control samples for each gene will be measured with laser scanning. A pre-processing procedure is then used to obtain expression difference values for each gene. In this example, the result will be 6 relative expression values (e.g., $\text{Log}_2(\textit{Post} / \textit{Pre})$) for each gene represented on the arrays.

Paired Design, Six Arrays



Null and Alternative Hypotheses

The paired test null hypothesis for each gene is H_0: $\mu_{diff} = \mu_0$, where μ_{diff} is the true mean difference in expression for a particular gene, and μ_0 is the hypothesized difference in expression. The alternative hypothesis may be any one of the following: H_a: $\mu_{diff} < \mu_0$, H_a: $\mu_{diff} > \mu_0$, or H_a: $\mu_{diff} \neq \mu_0$. The choice of the alternative hypothesis depends upon the goals of the research. For example, if the goal of the experiment is to determine which genes are differentially expressed, without regard to direction, the alternative hypothesis would be H_a: $\mu_{diff} \neq \mu_0$. If, however, the goal is to identify only genes which have increased expression after the treatment is applied, the alternative hypothesis would be H_a: $\mu_{diff} > \mu_0$. A common value for μ_0 in a paired sample experiment is 0.

T-Test Formula

For testing the hypothesis H_0: $\mu_{diff} = \mu_0$, the formula for the paired T-statistic is:

$$T_{paired} = \frac{\bar{x}_{paired} - \mu_o}{\frac{s_{paired}}{\sqrt{n}}}$$

where $\bar{x}_{paired}$ is mean difference in expression of n replicates for a given gene, μ_0 is the hypothesized mean, and s_{paired} is standard deviation of the differences of the n replicates. If the assumptions (described below) of the test are met, the distribution of T_{paired} is the standard t distribution with $n – 1$ degrees of freedom. P-values are obtained from T_{paired} by finding the proportion of the t distribution that is more extreme than T_{paired} .

Assumptions

The assumptions of the paired t-test are:

1. The data are continuous (not discrete). Because of the large range of possible intensities, microarray expression values can be considered continuous.
2. The data, i.e., the expression differences for the pairs, follow a normal probability distribution. This assumption can be examined in microarray data only if the number of arrays in the experiment is reasonably large (>100).
3. The sample of pairs is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. If the samples used in the microarray experiment are not random, bias may easily be introduced into the results.

One-Sample Design

The one-sample design is the simplest of all designs. A single mRNA or cDNA sample is obtained from each experimental unit of a single group. Each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of the group. The goal is to determine for each gene whether there is evidence that the expression is different from some null value. This design may be useful for determining whether or not each gene is expressed at all, or for comparing expression of each gene to a hypothesized expression level.

Null and Alternative Hypotheses

The one-sample null hypothesis for each gene is H_0: $\mu = \mu_0$, where μ is the true mean expression for that particular gene, and μ_0 is the hypothesized mean, or the mean to be compared against. The alternative hypothesis may be any one of the following: H_a: $\mu < \mu_0$, H_a: $\mu > \mu_0$, or H_a: $\mu \neq \mu_0$. The choice of the alternative hypothesis depends upon the goals of the research. For example, if the goal of the experiment is to determine which genes are expressed above a certain level, the alternative hypothesis would be H_a: $\mu > \mu_0$.

T-Test Formula

For testing the hypothesis H_0: $\mu = \mu_0$, the formula for the one-sample T-statistic is:

$$T = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}}$$

where $\overline{x}$ is mean expression of n replicates for a given gene, μ_0 is the hypothesized mean, and s is standard deviation of the n replicates. If the assumptions (described below) of the test are met, the distribution of T is the standard t distribution with $n – 1$ degrees of freedom. P-values are obtained from T by finding the proportion of the t distribution that is more extreme than T.

Assumptions

The assumptions of the one-sample T-test are:

1. The data are continuous (not discrete). Because of the large range of possible intensities, microarray expression values can be considered continuous.
2. The data follow the normal probability distribution. This assumption can be examined in microarray data only if the number of arrays in the experiment is reasonably large (>300).
3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample. If the samples used in the microarray experiment are not random, bias may easily be introduced into the results.

Technical Details

Multiple Testing Adjustment

When a one-sample/paired T-test is run for a replicated microarray experiment, the result is a list of P-values (Probability Level) that reflect the evidence of difference in expression. When hundreds or thousands of genes are investigated at the same time, many 'small' P-values will occur by chance, due to the natural variability of the process. It is therefore requisite to make an appropriate adjustment to the P-value (Probability Level), such that the likelihood of a false conclusion is controlled.

False Discovery Rate Table

The following table (adapted to the subject of microarray data) is found in Benjamini and Hochberg's (1995) false discovery rate article. In the table, m is the total number of tests, m_0 is the number of tests for which there is no difference in expression, R is the number of tests for which a difference is declared, and U, V, T, and S are defined by the combination of the declaration of the test and whether or not a difference exists, in truth.

	Declared Not Different	Declared Different	Total
A true difference in expression does not exist	U	V	m_0
There exists a true difference in expression	T	S	$m - m_0$
Total	$m - R$	R	m

In the table, the m is the total number of hypotheses tested (or total number of genes) and is assumed to be known in advance. Of the m null hypotheses tested, m_0 is the number of tests for which there is no difference in expression, R is the number of tests for which a difference is declared, and U, V, T, and S are defined by the combination of the declaration of the test and whether or not a difference exists, in truth. The random variables U, V, T, and S are unobservable.

Need for Multiple Testing Adjustment

Following the calculation of a raw P-value (Probability Level) for each test, P-value adjustments need be made to account in some way for multiplicity of tests. It is desirable that these adjustments minimize the number of genes that are falsely declared different (*V*) while maximizing the number of genes that are correctly declared different (*S*). To address this issue the researcher must know the comparative value of finding a gene to the price of a false positive. If a false positive is very expensive, a method that focuses on minimizing V should be employed. If the value of finding a gene is much higher than the cost of additional false positives, a method that focuses on maximizing *S* should be used.

Error Rates – P-Value Adjustment Techniques

Below is a brief description of three common error rates that are used for control of false positive declarations. The commonly used P-value adjustment technique for controlling each error rate is also described.

Per-Comparison Error Rate (PCER) – No Multiple Testing Adjustment

The per-comparison error rate (PCER) is defined as

$$\text{PCER} = E(V)/m\,,$$

where *E*(*V*) is the expected number of genes that are falsely declared different, and *m* is the total number of tests. Preserving the PCER is tantamount to ignoring multiple testing altogether. If a method is used which controls a PCER of 0.05 for 1,000 tests, approximately 50 out of 1,000 tests will falsely be declared significant. Using a method that controls the PCER will produce a list of genes that includes most of the genes for which there exists a true difference in expression (i.e., maximizes *S*), but it will also include a very large number of genes which are falsely declared to have a true difference in expression (i.e., does not appropriately minimize *V*). Controlling the PCER should be viewed as overly weak control of Type I error.

To obtain P-values (Probability Levels) that control the PCER, no adjustment is made to the P-value. To determine significance, the P-value is simply compared to the designated alpha.

Experiment-Wise Error Rate (EWER)

The experiment-wise error rate (EWER) is defined as

$$\text{EWER} = \Pr(V > 0),$$

where *V* is the number of genes that are falsely declared different. Controlling EWER is controlling the probability that a single null hypothesis is falsely rejected. If a method is used which controls a EWER of 0.05 for 1,000 tests, the probability that any of the 1,000 tests (collectively) is falsely rejected is 0.05. Using a method that controls the EWER will produce a list of genes that includes a small (depending also on sample size) number of the genes for which there exists a true difference in expression (i.e., limits *S*, unless the sample size is very large). However, the list of genes will include very few or no genes that are falsely declared to have a true difference in expression (i.e., stringently minimizes *V*). Controlling the EWER should be considered very strong control of Type I error.

Assuming the tests are independent, the well-known Bonferroni P-value adjustment produces adjusted P-values (Probability Levels) for which the EWER is controlled. The Bonferroni adjustment is applied to all *m* unadjusted *P*-values (p_j) as

$$\tilde{p}_j = \min(mp_j, 1)\,.$$

That is, each P-value (Probability Level) is multiplied by the number of tests, and if the result is greater than one, it is set to the maximum possible P-value of one.

False Discovery Rate (FDR)

The false discovery rate (FDR) (Benjamini and Hochberg, 1995) is defined as

$$\text{FDR} = E(\frac{V}{R}1_{\{R>0\}}) = E(\frac{V}{R} \mid R>0)\,\Pr(R>0),$$

where R is the number of genes that are declared significantly different, and V is the number of genes that are falsely declared different. Controlling FDR is controlling the expected *proportion* of falsely declared differences (false discoveries) to declared differences (true and false discoveries, together). If a method is used which controls a FDR of 0.05 for 1,000 tests, and 40 genes are declared different, it is expected that 40*0.05 = 2 of the 40 declarations are false declarations (false discoveries). Using a method that controls the FDR will produce a list of genes that includes an intermediate (depending also on sample size) number of genes for which there exists a true difference in expression (i.e., moderate to large S). However, the list of genes will include a small number of genes that are falsely declared to have a true difference in expression (i.e., moderately minimizes V). Controlling the FDR should be considered intermediate control of Type I error.

Multiple Testing Adjustment Comparison

The following table gives a summary of the multiple testing adjustment procedures and error rate control. The power to detect differences also depends heavily on sample size.

Common Adjustment Technique	Error Rate Controlled	Control of Type I Error	Power to Detect Differences
None	PCER	Minimal	High
Bonferroni	EWER	Strict	Low
Benjamini and Hochberg	FDR	Moderate	Moderate/High

Type I Error: Rejection of a null hypothesis that is true.

Calculating Power

When the standard deviation is unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which *Diff* > *0*. Additional details of calculating power in the one-sample/paired scenario are found in the PASS chapter for One Mean.

1. Find t_α such that $1 - T_{df}\left(t_\alpha\right) = \alpha$, where $T_{df}\left(t_\alpha\right)$ is the area under a central-*t* curve to the left of *x* and *df* = *n* - *1*.

2. Calculate: $x_a = t_\alpha \frac{\sigma}{\sqrt{n}}$.

3. Calculate the non-centrality parameter: $\lambda_a = \dfrac{Diff}{\frac{\sigma}{\sqrt{n}}}$.

4. Calculate: $t_a = \dfrac{x_a - Diff}{\frac{\sigma}{\sqrt{n}}} + \lambda_a$.

5. Calculate: Power = $1 - T'_{df,\lambda}\left(t_a\right)$, where $T'_{df,\lambda}\left(x\right)$ is the area under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ to the left of *x*.

Adjusting Alpha

Experiment-wise Error Rate

When the Bonferroni method will be used to control the experiment-wise error rate, α_{EWER}, of all tests, the adjusted α, α_{ADJ}, for each test is given by

$$\alpha_{ADJ} = \frac{\alpha_{EWER}}{Number\ of\ Tests}$$

α_{ADJ} is the value that is used in the power and sample size calculations.

False Discovery Rate

When a false discovery rate controlling method will be used to control the false discovery rate for the experiment, *fdr*, the adjusted α, α_{ADJ}, for each test is given by Jung (2005) and Chow, Shao, and Wang (2008):

$$\alpha_{ADJ} = \frac{(K)(1-\beta)(fdr)}{(N_T - K)(1 - fdr)}$$

where K is the number of genes with differential expression, β is the probability of a Type II error (not declaring a gene significant when it is), and N_T is the total number of tests.

α_{ADJ} is the value that is used in the power and sample size calculations. Because α_{ADJ} depends on β, α_{ADJ} must be solved iteratively when the calculation of power is desired.

Procedure Options

This section describes the options that are unique to this procedure. These are located on the panels associated with the Data, Options, and Reports tabs. To find out more about using the other tabs such as Plot Text, Axes, and Template, turn to the chapter entitled Procedure Templates at the beginning of this manual.

Data Tab

The Data tab contains most of the parameters and options involved in the power and sample size calculations.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Power and Beta* or *N*.

Select *N* when you want to determine the sample size needed to achieve a given power and false discovery rate (or alpha) error level.

Select *Power and Beta* when you want to calculate the power of an experiment.

Error Rates

Power for each Gene

Power is the probability of rejecting each null hypothesis when it is false. Power is equal to 1-Beta.

The POWER for each gene represents that probability of detecting differential expression when it exists.

RANGE: The valid range is from 0 to 1.

RECOMMENDED: Popular values for power are 0.8 and 0.9.

NOTES: You can enter a range of values such as *.70 .80 .90* or *.70 to .95 by .05*.

False Discovery (Alpha) Method

A type I error is declaring a gene to be differentially expressed when it is not. The two most common methods for controlling type I error in microarray expression studies are false discovery rate (FDR) control and Experiment-wise Error Rate (EWER) control.

- **FDR**

  Controlling the false discovery rate (FDR) controls the PROPORTION of genes that are falsely declared differentially expressed. For example, suppose that an FDR of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, 5 of the 100 genes are expected to be false discoveries.

- **EWER**

  Controlling the experiment-wise error rate (EWER) controls the PROBABILITY of ANY false declarations of differential expression, across all tests. For example, suppose that an EWER of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, the probability that even one of the 100 declarations is false is 0.05.

Recommendation: For exploratory studies where a list of candidate genes for further study is the goal, FDR is the recommended Type I error control method, because of its higher power.

For confirmatory studies where final determination of differential expression is the goal, EWER is the recommended Type I error control method, because of its strict control of false discoveries.

FDR or EWER Value

Specify the value for the False Discovery (Alpha) Method selected above.

RANGE: These levels are bounded by 0 and 1. Commonly, the chosen level is between 0.001 and 0.250

RECOMMENDED: FDR or EWER is often set to 0.05 for two-sided tests and to 0.025 for one-sided tests.

NOTE: You can enter a list of values such as *.05 .10 .15* or *.05 to .15 by .01*.

Sample Size

N (Number of Arrays)

Enter a value for the sample size (N). This is the number of arrays in the experiment. For two-channel paired experiments, this is the number of arrays, not the number of samples. You may enter a range such as *10 to 100 by 10* or a list of values separated by commas or blanks.

Effect Size

D (Minimum Mean Difference Detected)

Specify the true mean difference in expression (D) such that genes with true mean difference above D will be detected at the given power and corresponding sample size.

In paired expression studies, it is very common that the difference in expression is measured on the log scale (e.g., log2(A) – log2(B)). Values of D should reflect the differences that will be used in testing. For example, if log2 differences are used, D = 1 implies a two-fold difference in expression, while D = 2 implies a four-fold difference in expression.

When D is large, the resulting sample size will only detect the genes with extreme differential expression.

When D is small, a larger sample size is required to have power sufficient to detect these small differences in expression.

You can enter a range of values such as *1 2 3* or *0.2 to 2 by 0.1*.

S (Standard Deviation of Paired Differences in Expression)

Specify the standard deviation of paired differences. This standard deviation is assumed for all tests.

S should be on the same scale as D.

To obtain the standard deviation of paired differences from the standard deviation of expression, use SDpaired = (sqrt(2))*(SDexpression).

Because the true variation in paired differences will vary from gene to gene, it is recommended that a range of values be entered here.

You can enter a range of values such as *1 2 3 4 5* or *0.2 to 2 by 0.1*.

Number of Genes

Number of Genes Tested

Specify the number of genes for which hypothesis tests will be made.

This number will usually be the number of genes summarized on each array minus the number of housekeeping genes.

Only one number may be entered in this box.

Number of Genes – FDR Only

K (Number of Genes with Mean Difference > D)

Specify the number of genes for which a true mean difference in expression greater than D is expected.

K for EWER

The choice of K has no direct effect on the calculation of power or sample size when the False Discovery (Alpha) Method is set to EWER. K is not used when False Discovery (Alpha) method is set to EWER.

K for FDR

The choice of K has direct effect on the calculation of power or sample size when the False Discovery (Alpha) Method is set to FDR.

You can enter a range of values such as *10 20 30 40 50* or *20 to 100 by 10*.

Test

Alternative

Specify whether the hypothesis test for each gene is one-sided (directional) or two-sided (non-directional).

Recommendation: In most paired experiments, differential expression in either direction (up-regulation or down-regulation) is of interest. Such experiments should have the Two-Sided alternative hypothesis.

For experiments for determining whether there is expression above some threshold, a One-Sided alternative hypothesis is recommended. Often regulations dictate that the FDR or EWER level be divided by 2 for One-Sided alternative tests.

Test Type

Select the type of test that will be used when the analysis of the gene expression data is carried out.

- **T**

  The T-Test assumes the paired differences come from a normal distribution with UNKNOWN standard deviation (i.e., a standard deviation that will be estimated from the data).

- **Z**

  The Z-Test assumes the paired differences come from a normal distribution with KNOWN standard deviation.

Recommendation: Because it very rare to know the true standard deviation of paired differences in advance, T is the recommended test statistic.

Options Tab

The Options tab contains convergence and iteration options that are rarely changed.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank.

RECOMMENDED: 500 (or more)

Convergence Options

FDR Power Convergence

When FDR is selected for False Discovery (Alpha) Method, and Find (Solve For) is set to Power, the corresponding search algorithm will converge when the search criteria is below this value.

This value will rarely be changed from the default value.

RECOMMENDED: 0.0000000001

Example 1 – Finding Power

This example examines the power to detect differential expression for an experiment that involved 22 two-channel arrays. Two samples were obtained from each of 22 individuals. One of the two samples was randomly assigned the treatment and the other remained as the control. Following treatment, the two samples were exposed to a single microarray. Each microarray produced intensity information for 10,000 genes. The 22 arrays were pre-processed by subtracting the control intensity (Log2) from the treatment intensity for each gene on each array. Thus, a positive value implies upward expression in the treatment, while a negative value implies down-regulation in the treatment. In this example, the paired T-test was used to determine which genes were differentially expressed (upward or downward) following exposure to the treatment.

The researchers found very few differentially expressed genes, and wish to examine the power of the experiment to detect two-fold differential expression (Log2-scale difference of 1). Typical standard deviations of the Log2 paired differences ranged from 0.2 to 2.0.

The researchers guess the number of genes with at least 2-fold differential expression to be around 50, but will examine the effect of this estimate on power by trying 10 and 100 genes as well. A false discovery rate of 0.05 was used.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Sample or Paired T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **One-Sample or Paired Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power for each Gene	*Ignored since this is the Find setting*
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.05**
N (Number of Arrays)	**22**
D (Difference)	**1.0**
S (Standard Deviation)	**0.2 to 2 by .2**
Number of Genes Tested	**10000**
K (Number > D)	**10 50 100**
Alternative Hypothesis	**Two-Sided**
Test Type	**T**
Reports Tab	
Numeric Reports	**All Checked**
Number of Summary Statements	**1**
Show Plots	**Checked**
Interactive Format	**Unchecked**
Show Beta as Power	**Checked**

Annotated Output

Click the Run button to perform the calculations and generate the following output. The calculations should take a few moments.

Numeric Results

```
Numeric Results for One-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
Number of Genes Tested: 10000

                           Minimum      Std.                Number      False
               Number         Mean      Dev.     Effect      Genes  Discovery
            of Arrays   Difference  of Diff.       Size  To Detect       Rate
Power             (N)          (D)       (S)       (ES)        (K)      (FDR)       Beta
1.00000            22          1.0       0.2      5.000         10     0.0500    0.00000
1.00000            22          1.0       0.2      5.000         50     0.0500    0.00000
1.00000            22          1.0       0.2      5.000        100     0.0500    0.00000
1.00000            22          1.0       0.4      2.500         10     0.0500    0.00000
1.00000            22          1.0       0.4      2.500         50     0.0500    0.00000
1.00000            22          1.0       0.4      2.500        100     0.0500    0.00000
0.98617            22          1.0       0.6      1.667         10     0.0500    0.01383
0.99793            22          1.0       0.6      1.667         50     0.0500    0.00207
0.99924            22          1.0       0.6      1.667        100     0.0500    0.00076
0.71696            22          1.0       0.8      1.250         10     0.0500    0.28304
0.89092            22          1.0       0.8      1.250         50     0.0500    0.10908
0.93538            22          1.0       0.8      1.250        100     0.0500    0.06462
.                   .            .         .          .          .          .          .
.                   .            .         .          .          .          .          .
.                   .            .         .          .          .          .          .

References
Jung, S.-H. 2005. Sample size for FDR-control in microarray data analysis. Bioinformatics: Vol. 21 no. 14, pp.
3097-3104. Oxford University Press.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd
Edition. Blackwell Science. Malden, MA.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions
Power is the individual probability of detecting each gene with true mean difference > D.
N is the number of arrays required to achieve the corresponding power.
D is the smallest difference in expression for which this power and sample size are valid.
S is the standard deviation estimate for the paired differences used in each test.
ES, or D/S, is the relative magnitude of the true mean expression difference for the genes with true mean
difference > D.
K is the number of genes with true mean difference > D.
FDR is the expected proportion of false declarations of differential expression to total declarations of
differential expression.
Beta is the individual probability of failing to detect each gene with true mean difference > D.

Summary Statements
A sample size of 22 achieves 100.00% power for each gene to detect a true difference in
expression of at least 1.0 with an estimated standard deviation of 0.2 with a false discovery
rate of 0.0500 using a two-sided one-sample T-Test. Of the 10 genes with anticipated true mean
difference in expression > 1.0, 9 are expected to be detected.
```

This report shows the values of each of the parameters, one scenario per row. The values of power and beta were calculated from the other parameters.

The definitions of each column are given in the Report Definitions section.

Additional Numeric Result Detail

Additional Numeric Result Detail for One-Sample T-Test
Number of Genes To Be Tested: 10000

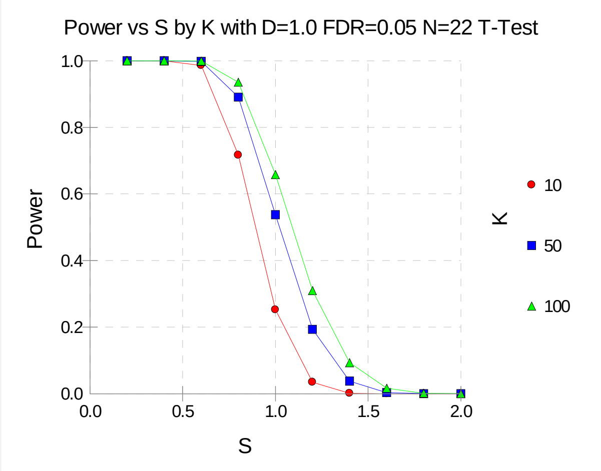
Power	Number of Arrays (N)	Minimum Mean Difference (D)	Std. Dev. of Diff. (S)	Number Genes To Detect (K)	False Discovery Rate (FDR)	Single Gene Alpha	Prob To Detect All K
1.00000	22	1.0	0.2	10	0.0500	0.0000527	1.00000
1.00000	22	1.0	0.2	50	0.0500	0.0002645	1.00000
1.00000	22	1.0	0.2	100	0.0500	0.0005316	1.00000
1.00000	22	1.0	0.4	10	0.0500	0.0000527	1.00000
1.00000	22	1.0	0.4	50	0.0500	0.0002645	1.00000
1.00000	22	1.0	0.4	100	0.0500	0.0005316	1.00000
0.98617	22	1.0	0.6	10	0.0500	0.0000520	0.86996
0.99793	22	1.0	0.6	50	0.0500	0.0002639	0.90158
0.99924	22	1.0	0.6	100	0.0500	0.0005312	0.92649
0.71696	22	1.0	0.8	10	0.0500	0.0000378	0.03589
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Report Definitions
Power is the individual probability of detecting each gene with true mean difference > D.
N is the number of arrays required to achieve the corresponding power.
D is the smallest difference in expression for which this power and sample size are valid.
S is the standard deviation estimate for the paired differences used in each test.
K is the number of genes with true mean difference > D.
FDR is the expected proportion of false declarations of differential expression to total declarations of differential expression.
Single Gene Alpha is the probability of falsely declaring differential expression for an individual gene.
Prob to Detect All K is the probability of declaring differential expression for all K genes that have true mean difference > D.

This report shows additionally the single gene alpha and the probability of detecting all K differentially expressed genes.

The definitions of each column are given in the Report Definitions section.

Plots Section



This plot shows the relationship between power and the standard deviation of the differences for various three values of K.

Example 2 – Finding the Sample Size

This example determines the number of two-channel arrays needed to achieve 80% power to detect differential expression for each gene. Two samples will be obtained from each of the sampled individuals. One of the two samples will be randomly assigned the treatment and the other will remain as the control. Following treatment, the two samples will be exposed to a single microarray. Each microarray will produce intensity information for 12,682 genes. The arrays will be pre-processed by subtracting the control intensity (Log2) from the treatment intensity for each gene on each array. Thus, a positive value implies upward expression in the treatment, while a negative value implies down-regulation in the treatment. The paired T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment.

The researchers wish to detect differential expression that is two-fold or greater (Log2-scale difference of 1). Typical standard deviations of the Log2 paired differences for this experiment are expected to range from 0.2 to 2.0.

The researchers guess the number of genes with at least 2-fold differential expression to be around 50, but will examine the effect of this estimate on sample size by trying 10 and 100 genes as well. A false discovery rate of 0.05 will be used.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Sample or Paired T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **One-Sample or Paired Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power for each Gene	**0.20**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.05**
N (Number of Arrays)	*Ignored since this is the Find setting*
D (Difference)	**1**
S (Standard Deviation)	**.2 to 2 by .2**
Number of Genes Tested	**12682**
K (Number > D)	**10 50 100**
Alternative Hypothesis	**Two-Sided**
Test Type	**T**
Reports Tab	
Numeric Reports	**All Checked but Numeric Report Detail**
Number of Summary Statements	**1**
Show Plots	**Checked**
Interactive Format	**Unchecked**
Show Beta as Power	**Checked**

Output

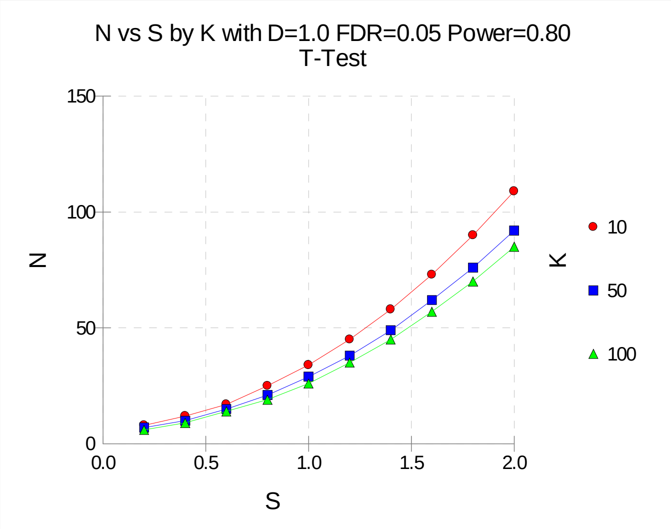
Click the Run button to perform the calculations and generate the following output. The calculations may take a few moments.

Numeric Results

```
Numeric Results for One-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
Number of Genes Tested: 12682

                           Minimum        Std.                  Number        False
                Number        Mean        Dev.     Effect        Genes    Discovery
             of Arrays  Difference    of Diff.       Size    To Detect         Rate
Power              (N)         (D)         (S)       (ES)          (K)        (FDR)       Beta
0.96741              8         1.0         0.2      5.000           10       0.0500    0.03259
0.97509              7         1.0         0.2      5.000           50       0.0500    0.02491
0.91190              6         1.0         0.2      5.000          100       0.0500    0.08810
0.88530             12         1.0         0.4      2.500           10       0.0500    0.11470
0.86231             10         1.0         0.4      2.500           50       0.0500    0.13769
0.83278              9         1.0         0.4      2.500          100       0.0500    0.16722
0.81531             17         1.0         0.6      1.667           10       0.0500    0.18469
0.85472             15         1.0         0.6      1.667           50       0.0500    0.14528
0.86398             14         1.0         0.6      1.667          100       0.0500    0.13602
0.83661             25         1.0         0.8      1.250           10       0.0500    0.16339
0.82805             21         1.0         0.8      1.250           50       0.0500    0.17195
.                    .           .           .          .            .            .          .
.                    .           .           .          .            .            .          .
.                    .           .           .          .            .            .          .
```

This report shows the values of each of the parameters, one scenario per row. The sample size (number of arrays) estimates were calculated from the other parameters. The power is the actual power produced by the given sample size.

Plots Section



This plot shows the relationship between sample size and the standard deviation of the differences for three values of K.

Example 3 – Finding the Minimum Detectable Difference

This example finds the minimum difference in expression that can be detected with 90% power from a microarray experiment with 14 two-channel arrays. The 14 arrays permit tests on 5,438 genes. The arrays will be pre-processed by subtracting the control intensity (Log2) from the treatment intensity for each gene on each array. Thus, a positive value implies upward expression in the treatment, while a negative value implies down-regulation in the treatment. The paired T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment. Standard deviations of the Log2 paired differences for this experiment range from 0.2 to 1.8.

In this example we will examine a range for K (the number of genes with mean difference greater than the minimum detectable difference), since this should vary with the mean difference chosen. A false discovery rate of 0.05 will be used.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Sample or Paired T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **One-Sample or Paired Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Mean Difference**
Power for each Gene	**0.90**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.05**
N (Number of Arrays)	**14**
D (Difference)	*Ignored since this is the Find setting*
S (Standard Deviation)	**.2 to 1.8 by .4**
Number of Genes Tested	**5438**
K (Number > D)	**10 to 50 by 10**
Alternative Hypothesis	**Two-Sided**
Test Type	**T**
Reports Tab	
Numeric Reports	**All Checked but Numeric Report Detail**
Number of Summary Statements	**1**
Show Plots	**Checked**
Interactive Format	**Unchecked**
Show Beta as Power	**Checked**
Mean Difference and SD Decimals	**4**

Output

Click the Run button to perform the calculations and generate the following output. The calculations may take a few moments.

Numeric Results

Numeric Results for One-Sample T-Test

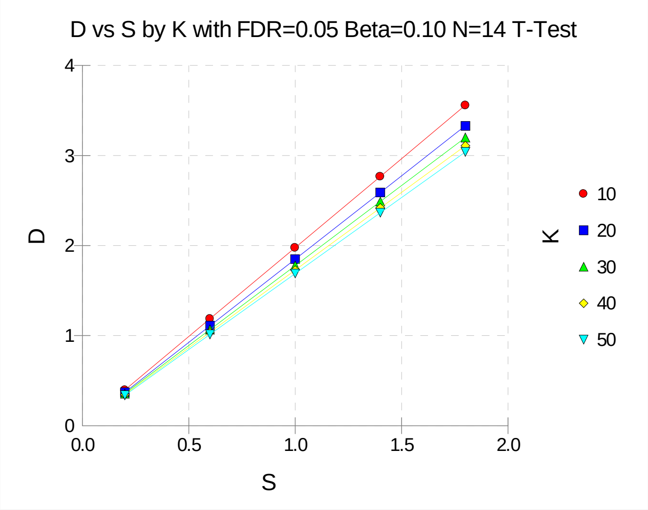
Null Hypothesis: MeanDiff = 0   Alternative Hypothesis: MeanDiff <> 0

Number of Genes Tested: 5438

		Minimum	Std.		Number	False	
	Number	Mean	Dev.	Effect	Genes	Discovery	
	of Arrays	Difference	of Diff.	Size	To Detect	Rate	
Power	**(N)**	**(D)**	**(S)**	**(ES)**	**(K)**	**(FDR)**	**Beta**
0.90000	14	0.3951	0.2000	1.976	10	0.0500	0.10000
0.90000	14	0.3699	0.2000	1.849	20	0.0500	0.10000
0.90000	14	0.3555	0.2000	1.777	30	0.0500	0.10000
0.90000	14	0.3454	0.2000	1.727	40	0.0500	0.10000
0.90000	14	0.3377	0.2000	1.689	50	0.0500	0.10000
0.90000	14	1.1854	0.6000	1.976	10	0.0500	0.10000
0.90000	14	1.1096	0.6000	1.849	20	0.0500	0.10000
0.90000	14	1.0664	0.6000	1.777	30	0.0500	0.10000
0.90000	14	1.0363	0.6000	1.727	40	0.0500	0.10000
0.90000	14	1.0132	0.6000	1.689	50	0.0500	0.10000
0.90000	14	1.9756	1.0000	1.976	10	0.0500	0.10000
0.90000	14	1.8493	1.0000	1.849	20	0.0500	0.10000
0.90000	14	1.7774	1.0000	1.777	30	0.0500	0.10000
0.90000	14	1.7272	1.0000	1.727	40	0.0500	0.10000
0.90000	14	1.6887	1.0000	1.689	50	0.0500	0.10000
0.90000	14	2.7658	1.4000	1.976	10	0.0500	0.10000
0.90000	14	2.5890	1.4000	1.849	20	0.0500	0.10000
0.90000	14	2.4884	1.4000	1.777	30	0.0500	0.10000
0.90000	14	2.4181	1.4000	1.727	40	0.0500	0.10000
0.90000	14	2.3642	1.4000	1.689	50	0.0500	0.10000
0.90000	14	3.5561	1.8000	1.976	10	0.0500	0.10000
0.90000	14	3.3287	1.8000	1.849	20	0.0500	0.10000
0.90000	14	3.1993	1.8000	1.777	30	0.0500	0.10000
0.90000	14	3.1090	1.8000	1.727	40	0.0500	0.10000
0.90000	14	3.0397	1.8000	1.689	50	0.0500	0.10000

This report shows the values of each of the parameters, one scenario per row. The Minimum Mean Difference (D) estimates were calculated from the other parameters.

Plots Section



This plot shows the relationship between D (the minimum detectable difference on the Log2 scale) and the standard deviation of the differences for five values of K.

Example 4 – Validation (EWER) using Stekel

Stekel (2003), pp. 226-228, gives an example in which N = 20, D = 1, and S = 0.68 for a two-sided paired T-Test. The number of genes tested is 6500. The control of false discoveries is “no more than one false positive.” This corresponds to an EWER value of 0.975. The power obtained for this example is 0.94.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Sample or Paired T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **One-Sample or Paired Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power for each Gene	*Ignored since this is the Find setting*
False Discovery (Alpha) Method	**EWER (Experiment-wise Error Rate)**
FDR or EWER Value	**0.975**
N (Number of Arrays)	**20**
D (Difference)	**1**
S (Standard Deviation)	**0.68**
Number of Genes Tested	**6500**
K (Number > D)	*Ignored since EWER is used*
Alternative Hypothesis	**Two-Sided**
Test Type	**T**
Reports Tab	
Reports	**All Checked but Numeric Report Detail**
Mean Difference and SD Decimals	**2**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
Number of Genes Tested: 6500

                              Minimum       Std.               Experiment
                 Number          Mean       Dev.      Effect        -Wise        Single
               of Arrays   Difference    of Diff.       Size   Error Rate          Gene
Power                (N)          (D)        (S)        (ES)      (EWER)         Alpha         Beta
0.93591               20         1.00       0.68       1.471      0.9750     0.0001500      0.06409
```

The power of 0.93591 matches Stekel’s result.

Example 5 – Validation (EWER) using Lee

Lee (2004), pp. 218-220, gives an example in which Power = 0.90, D = 1.0 1.5 2.0 2.5 and S = 1.0 for a two-sided paired Z-Test. The number of genes tested is 1000. The control of false discoveries is 0.5. This corresponds to an EWER value of 0.5. This setup corresponds to the upper left corner of Table 14.3 on page 219. The sample sizes obtained for this setup are 23, 11, 6, and 4, respectively.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Sample or Paired T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **One-Sample or Paired Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power for each Gene	**0.90**
False Discovery (Alpha) Method	**EWER (Experiment-wise Error Rate)**
FDR or EWER Value	**0.5**
N (Number of Arrays)	*Ignored since this is the Find setting*
D (Difference)	**1.0 1.5 2.0 2.5**
S (Standard Deviation)	**1.0**
Number of Genes Tested	**1000**
K (Number > D)	*Ignored since EWER is used*
Alternative Hypothesis	**Two-Sided**
Test Type	**Z**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample Z-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
Number of Genes Tested: 1000

                          Minimum         Std.               Experiment
             Number          Mean         Dev.      Effect        -Wise       Single
          of Arrays    Difference     of Diff.        Size   Error Rate         Gene
Power           (N)           (D)          (S)        (ES)       (EWER)        Alpha         Beta
0.90576          23          1.00         1.00       1.000       0.5000    0.0005000      0.09424
0.93244          11          1.50         1.00       1.500       0.5000    0.0005000      0.06756
0.92194           6          2.00         1.00       2.000       0.5000    0.0005000      0.07806
0.93565           4          2.50         1.00       2.500       0.5000    0.0005000      0.06435
```

Sample sizes of 23, 11, 6, and 4 match the results shown in Lee (2004).

Example 6 – Validation (FDR) using Jung

Jung (2005), page 3100, gives an example for the sample size needed to control FDR in a two-sample Z-Test. This example is repeated in Chow, Shao, and Wang (2008). We adapt the effect size in this validation to correspond to a one-sample test. Namely, the effect size is reduced by one half. In the example, Power = 0.60 (from 24/40), D = 1.0, and S = 1.0 for a one-sided two-sample Z-Test. We use S = 2.0 to correspond to the equivalent in the one-sample test. The number of genes tested is 4000. The FDR level is 1%. This setup corresponds to Example 1 on page 3100. The required sample size obtained for this setup is 68.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **One-Sample or Paired T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **One-Sample or Paired Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N (Sample Size)**
Power for each Gene	**0.60**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.01**
N (Number of Arrays)	*Ignored since this is the Find setting*
D (Difference)	**1.0**
S (Standard Deviation)	**2.0**
Number of Genes Tested	**4000**
K (Number > D)	**40**
Alternative Hypothesis	**One-Sided**
Test Type	**Z**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for One-Sample Z-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff > 0
Number of Genes Tested: 4000

                           Minimum      Std.                 Number       False
                Number        Mean      Dev.     Effect       Genes   Discovery
             of Arrays  Difference   of Diff.      Size   To Detect        Rate
Power              (N)         (D)       (S)       (ES)         (K)       (FDR)        Beta
0.61099             68        0.50      1.00      0.500          40      0.0100     0.38901
```

A sample size of 68 matches the result shown in Jung (2005). For Example 3 in Jung (2005), the alternative hypothesis is two-sided and results in a sample size of 73. This result may be validated in ***PASS*** by changing Alternative to Two-Sided in this example.

Chapter 615

Two-Sample T-Test for Microarray Data

Introduction

This chapter describes how to estimate power and sample size (number of arrays) for 2 group (two-sample) microarray studies using the PASS: Two-Sample T-Test for Microarray Data procedure. False discovery rate and experiment-wise error rate control methods are available in this procedure. Values that can be varied in this procedure are power, false discovery rate and experiment-wise error rate, sample sizes (numbers of arrays) in each group, the minimum mean difference detected, the standard deviations in each group, and in the case of false discovery rate control, the number of genes with minimum mean difference.

Two-Sample Design

In a two-sample design, two groups are compared, which we will call Treatment 1 and Treatment 2. Several experimental units are randomly assigned to each of the two treatment groups. A single mRNA or cDNA sample is obtained from each experimental unit of both groups. Each sample is exposed to a single microarray, resulting in a single expression value for each gene for each unit of each treatment group. The goal is to determine for each gene whether there is evidence that the expression is different between the two groups.

Null and Alternative Hypotheses

The two-sample null and alternative hypotheses are described here in terms of treatment groups: Treatment 1 and Treatment 2. These groups could equally be labeled Treatment A and Treatment B, Control and Treatment, etc. The two-sample null hypothesis for each gene is H_0: $\mu_1 = \mu_2$, where μ_1 is the true mean expression for that particular gene in the Treatment 1 environment, and μ_2 is the true mean expression for that particular gene following Treatment 2. The alternative hypothesis may be any one of the following: H_a: $\mu_1 < \mu_2$, H_a: $\mu_1 > \mu_2$, or H_a: $\mu_1 \neq \mu_2$. The choice of the alternative hypothesis depends upon the goals of the research. For example, if the goal of the experiment is only to determine which genes are up-regulated (increase in expression) over Treatment 1 when Treatment 2 is imposed, the alternative hypothesis would be H_a: $\mu_1 > \mu_2$. If the goal instead is to determine which genes are differentially expressed (up-regulated or down-regulated) when compared to the other treatment, the alternative hypothesis is H_a: $\mu_1 \neq \mu_2$.

Assumptions

The following assumptions are made when using the two-sample T-test or the Mann-Whitney *U* test. One of the reasons for the popularity of the T-test is its robustness in the face of assumption violation. However, if an assumption is not met even approximately, the significance levels and the power of the T-test are unknown. You should take the appropriate steps to check the assumptions before you make important decisions based on these tests.

Two-Sample T-Test Assumptions

The assumptions of the two-sample T-test are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)
4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired T-test).
5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Mann-Whitney U Test Assumptions

The assumptions of the Mann-Whitney *U* test for difference in means are:

1. The variable of interest is continuous (not discrete). The measurement scale is at least ordinal.
2. The probability distributions of the two populations are identical, except for location. That is, the variances are equal.
3. The two samples are independent.
4. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Technical Details

Multiple Testing Adjustment

When the two-sample T-test is run for a replicated microarray experiment, the result is a list of P-values (Probability Levels) that reflect the evidence of difference in expression. When hundreds or thousands of genes are investigated at the same time, many ‘small’ P-values will occur by chance, due to the natural variability of the process. It is therefore requisite to make an appropriate adjustment to the P-value (Probability Level), such that the likelihood of a false conclusion is controlled.

Benjamini and Hochberg’s (1995) False Discovery Rate Table

The following table (adapted to the subject of microarray data) is found in Benjamini and Hochberg’s (1995) false discovery rate article. In the table, *m* is the total number of tests, m_0 is the number of tests for which there is no difference in expression, *R* is the number of tests for which a difference is declared, and *U*, *V*, *T*, and *S* are defined by the combination of the declaration of the test and whether or not a difference exists, in truth.

	Declared Not Different	Declared Different	Total
A true difference in expression does not exist	*U*	*V*	m_0
There exists a true difference in expression	*T*	*S*	$m - m_0$
Total	$m - R$	*R*	*m*

In the table, the *m* is the total number of hypotheses tested (or total number of genes) and is assumed to be known in advance. Of the *m* null hypotheses tested, m_0 is the number of tests for which there is no difference in expression, *R* is the number of tests for which a difference is declared, and *U*, *V*, *T*, and *S* are defined by the combination of the declaration of the test and whether or not a difference exists, in truth. The random variables *U*, *V*, *T*, and *S* are unobservable.

Need for Multiple Testing Adjustment

Following the calculation of a raw P-value (Probability Level) for each test, P-value adjustments need be made to account in some way for multiplicity of tests. It is desirable that these adjustments minimize the number of genes that are falsely declared different (*V*) while maximizing the number of genes that are correctly declared different (*S*). To address this issue the researcher must know the comparative value of finding a gene to the price of a false positive. If a false positive is very expensive, a method that focuses on minimizing V should be employed. If the value of finding a gene is much higher than the cost of additional false positives, a method that focuses on maximizing *S* should be used.

Error Rates – P-Value Adjustment Techniques

Below is a brief description of three common error rates that are used for control of false positive declarations. The commonly used P-value adjustment technique for controlling each error rate is also described.

Per-Comparison Error Rate (PCER) – No Multiple Testing Adjustment

The per-comparison error rate (PCER) is defined as

$$\text{PCER} = E(V)/m,$$

where *E*(*V*) is the expected number of genes that are falsely declared different, and *m* is the total number of tests. Preserving the PCER is tantamount to ignoring multiple testing altogether. If a method is used which controls a PCER of 0.05 for 1,000 tests, approximately 50 out of 1,000 tests will falsely be declared significant. Using a method that controls the PCER will produce a list of genes that includes most of the genes for which there exists a true difference in expression (i.e., maximizes *S*), but it will also include a very large number of genes which are falsely declared to have a true difference in expression (i.e., does not appropriately minimize *V*). Controlling the PCER should be viewed as overly weak control of Type I error.

To obtain P-values (Probability Levels) that control the PCER, no adjustment is made to the P-value. To determine significance, the P-value is simply compared to the designated alpha.

Experiment-Wise Error Rate (EWER)

The experiment-wise error rate (EWER) is defined as

$$\text{EWER} = \Pr(V > 0),$$

where *V* is the number of genes that are falsely declared different. Controlling EWER is controlling the probability that a single null hypothesis is falsely rejected. If a method is used which controls a EWER of 0.05 for 1,000 tests, the probability that any of the 1,000 tests (collectively) is falsely rejected is 0.05. Using a method that controls the EWER will produce a list of genes that includes a small (depending also on sample size) number of the genes for which there exists a true difference in expression (i.e., limits *S*, unless the sample size is very large). However, the list of genes will include very few or no genes that are falsely declared to have a true difference in expression (i.e., stringently minimizes *V*). Controlling the EWER should be considered very strong control of Type I error.

Assuming the tests are independent, the well-known Bonferroni P-value adjustment produces adjusted P-values (Probability Levels) for which the EWER is controlled. The Bonferroni adjustment is applied to all *m* unadjusted *P*-values (p_j) as

$$\tilde{p}_j = \min(mp_j, 1) .$$

That is, each P-value (Probability Level) is multiplied by the number of tests, and if the result is greater than one, it is set to the maximum possible P-value of one.

False Discovery Rate (FDR)

The false discovery rate (FDR) (Benjamini and Hochberg, 1995) is defined as

$$\text{FDR} = E(\frac{V}{R} 1_{\{R>0\}}) = E(\frac{V}{R} \mid R > 0) \Pr(R > 0) ,$$

where *R* is the number of genes that are declared significantly different, and *V* is the number of genes that are falsely declared different. Controlling FDR is controlling the expected *proportion* of falsely declared differences (false discoveries) to declared differences (true and false discoveries, together). If a method is used which controls a FDR of 0.05 for 1,000 tests, and 40 genes are declared different, it is expected that 40*0.05 = 2 of the 40 declarations are false declarations (false discoveries). Using a method that controls the FDR will produce a list of genes that includes an intermediate (depending also on sample size) number of genes for which there exists a true difference in expression (i.e., moderate to large S). However, the list of genes will include a small number of genes that are falsely declared to have a true difference in expression (i.e., moderately minimizes V). Controlling the FDR should be considered intermediate control of Type I error.

Assuming the tests are independent, the Benjamini and Hochberg P-value adjustment produces adjusted P-values (Probability Levels) for which the FDR is controlled. These adjusted *P*-values are found as

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m} \{\min(\frac{m}{k} p_{r_k}, 1)\} ,$$

where $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$ are the observed ordered unadjusted *P*-values. The procedure is defined in Benjamini and Hochberg (1995). The corresponding adjusted *P*-value definition given here is found in Dudoit, Shaffer, and Boldrick (2003).

Multiple Testing Adjustment Comparison

The following table gives a summary of the multiple testing adjustment procedures and error rate control. The power to detect differences also depends heavily on sample size.

Common Adjustment Technique	Error Rate Controlled	Control of Type I Error	Power to Detect Differences
None	PCER	Minimal	High
Bonferroni	EWER	Strict	Low
Benjamini and Hochberg	FDR	Moderate	Moderate/High

Type I Error: Rejection of a null hypothesis that is true.

Calculating Power

Additional details of calculating power in the two-group scenario are found in the PASS chapter for Two Means.

There are four separate situations, each requiring different formulas. Let the means of the two populations be represented by μ_1 and μ_2. The difference between these means will be represented by *d*. Let the standard deviations of the two populations be represented as σ_1 and σ_2.

Case 1 – Standard Deviations Known and Equal (Z)

When $\sigma_1 = \sigma_2 = \sigma$ and are known, the power of the *t* test is calculated as follows for a directional alternative (one-tailed test) in which $d > 0$.

1. Find z_α such that $1 - \Phi(z_\alpha) = \alpha$, where $\Phi(x)$ is the area under the standardized normal curve to the left of *x*.
2. Calculate: $\sigma_{\bar{x}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$
3. Calculate: $z_p = \frac{z_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}}$
4. Calculate: Power $= 1 - \Phi(z_p)$

Case 2 – Standard Deviations Known and Unequal (Z)

When $\sigma_1 \neq \sigma_2$ and are known, the power is calculated as follows for a directional alternative (one-tailed test) in which $d > 0$.

1. Find z_α such that $1 - \Phi(z_\alpha) = \alpha$, where $\Phi(x)$ is the area under the standardized normal curve to the left of *x*.
2. Calculate: $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$
3. Calculate: $z_p = \frac{z_\alpha \sigma_{\bar{x}} \text{ - } d}{\sigma_{\bar{x}}}$
4. Calculate: $\text{Power} = 1 - \Phi(z_p)$

Case 3 – Standard Deviations Unknown and Equal

When $\sigma_1 = \sigma_2 = \sigma$ and are unknown, the power of the T-Test is calculated as follows for a directional alternative (one-tailed test) in which $d > 0$.

1. Find t_α such that $1 - T_{df}(t_\alpha) = \alpha$, where $T_{df}(t_\alpha)$ is the area under a central-*t* curve to the left of *x* and $df = N_1 + N_2 - 2$.
2. Calculate: $\sigma_{\bar{x}} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$
3. Calculate the noncentrality parameter: $\lambda = \frac{d}{\sigma_{\bar{x}}}$
4. Calculate: $t_p = \frac{t_\alpha \sigma_{\bar{x}} \text{ - } d}{\sigma_{\bar{x}}} + \lambda$
5. Calculate: Power = $1 - T'_{df,\lambda}(t_p)$, where $T'_{df,\lambda}(x)$ is the area under a noncentral-*t* curve with degrees of freedom *df* and noncentrality parameter λ to the left of *x*.

Case 4 – Standard Deviations Unknown and Unequal

When $\sigma_1 \neq \sigma_2$ and are unknown, the power is calculated as follows for a directional alternative (one-tailed test) in which $d > 0$. Note that in this case, an approximate T-Test is used.

1. Calculate: $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$.

2. Calculate: $f = \dfrac{\sigma_{\bar{x}}^4}{\dfrac{\sigma_1^4}{N_1^2(N_1+1)} + \dfrac{\sigma_2^4}{N_2^2(N_2+1)}} - 2$

   which is the adjusted degrees of freedom. Often, this is rounded to the next highest integer.

3. Find t_α such that $1 - T_f(t_\alpha) = \alpha$, where $T_f(t_\alpha)$ is the area to the left of *x* under a central-*t* curve with *f* degrees of freedom.

4. Calculate: $\lambda = \dfrac{d}{\sigma_{\bar{x}}}$,1 the noncentrality parameter.

5. Calculate: $t_p = \dfrac{t_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}} + \lambda$

6. Calculate: Power = $1 - T'_{f,\lambda}(t_p)$, where $T'_{f,\lambda}(x)$ is the area to the left of *x* under a noncentral-*t* curve with degrees of freedom *f* and noncentrality parameter λ.

Adjusting Alpha

Experiment-wise Error Rate

When the Bonferroni method will be used to control the experiment-wise error rate, α_{EWER}, of all tests, the adjusted α, α_{ADJ}, for each test is given by

$$\alpha_{ADJ} = \frac{\alpha_{EWER}}{Number\ of\ Tests}$$

α_{ADJ} is the value that is used in the power and sample size calculations.

False Discovery Rate

When a false discovery rate controlling method will be used to control the false discovery rate for the experiment, *fdr*, the adjusted α, α_{ADJ}, for each test is given by Jung (2005) and Chow, Shao, and Wang (2008):

$$\alpha_{ADJ} = \frac{(K)(1-\beta)(fdr)}{(N_T - K)(1 - fdr)}$$

where *K* is the number of genes with differential expression, β is the probability of a Type II error (not declaring a gene significant when it is), and N_T is the total number of tests.

α_{ADJ} is the value that is used in the power and sample size calculations. Because α_{ADJ} depends on β, α_{ADJ} must be solved iteratively when the calculation of power is desired.

Procedure Options

This section describes the options that are unique to this procedure. These are located on the panels associated with the Data, Options, and Reports tabs. To find out more about using the other tabs such as Plot Text, Axes, and Template, turn to the chapter entitled Procedure Templates at the beginning of this manual.

Data Tab

The Data tab contains most of the parameters and options involved in the power and sample size calculations.

Solve For

Find (Solve For)

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Beta and Power* or *N1*.

Select *N1* when you want to determine the sample sizes needed in each group to achieve a given power and false discovery rate (or alpha) error level.

Select *Beta and Power* when you want to calculate the power of an experiment.

Error Rates

Power for each Gene

Power is the probability of rejecting each null hypothesis when it is false. Power is equal to 1-Beta.

The POWER for each gene represents that probability of detecting differential expression when it exists.

RANGE: The valid range is from 0 to 1.

RECOMMENDED: Popular values for power are 0.8 and 0.9.

NOTES: You can enter a range of values such as *.70 .80 .90* or *.70 to .95 by .05*.

False Discovery (Alpha) Method

A type I error is declaring a gene to be differentially expressed when it is not. The two most common methods for controlling type I error in microarray expression studies are false discovery rate (FDR) control and Experiment-wise Error Rate (EWER) control.

- **FDR**

  Controlling the false discovery rate (FDR) controls the PROPORTION of genes that are falsely declared differentially expressed. For example, suppose that an FDR of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, 5 of the 100 genes are expected to be false discoveries.

- **EWER**

  Controlling the experiment-wise error rate (EWER) controls the PROBABILITY of ANY false declarations of differential expression, across all tests. For example, suppose that an EWER of 0.05 is used for 10000 tests (on 10000 genes). If differential expression is declared for 100 of the 10000 genes, the probability that even one of the 100 declarations is false is 0.05.

Recommendation: For exploratory studies where a list of candidate genes for further study is the goal, FDR is the recommended Type I error control method, because of its higher power.

For confirmatory studies where final determination of differential expression is the goal, EWER is the recommended Type I error control method, because of its strict control of false discoveries.

FDR or EWER Value

Specify the value for the False Discovery (Alpha) Method selected above.

RANGE: These levels are bounded by 0 and 1. Commonly, the chosen level is between 0.001 and 0.250

RECOMMENDED: FDR or EWER is often set to 0.05 for two-sided tests and to 0.025 for one-sided tests.

NOTE: You can enter a list of values such as *.05 .10 .15* or *.05 to .15 by .01*.

Sample Size

N1 (Number of Arrays, Group 1)

Enter a value (or range of values) for the sample size of this group 1. Note that these values are ignored when you are solving for *N1*. You may enter a range of values such as *10 to 100 by 10*.

N2 (Number of Arrays, Group 2)

Enter a value (or range of values) for the sample size of group 2 or enter *Use R* to base *N2* on the value of *N1*. You may enter a range of values such as *10 to 100 by 10.*

- **Use R**

  When *Use R* is entered here, *N2* is calculated using the formula

  $$N2 = [R(N1)]$$

  where *R* is the Sample Allocation Ratio and the operator [*Y*] is the first integer greater than or equal to *Y*. For example, if you want *N1* = *N2*, select *Use R* and set *R* = 1.

R (Sample Allocation Ratio)

Enter a value (or range of values) for *R*, the allocation ratio between samples. This value is only used when *N2* is set to *Use R*.

When used, *N2* is calculated from *N1* using the formula: *N2* = [*R*(*N1*)] where [*Y*] is the next integer greater than or equal to *Y*. Note that setting *R* = 1.0 forces *N2* = *N1*.

Effect Size

D (Minimum Mean Difference Detected)

Specify the true mean difference in expression (D) such that genes with true mean difference above D will be detected at the given power and corresponding sample size.

In expression studies, it is very common that the expressions are measured on the log scale. Values of D should reflect the differences that will be used in testing. For example, if the log2 scale is used, D = 1 implies a two-fold difference in expression, while D = 2 implies a four-fold difference in expression.

When D is large, the resulting sample size will only detect the genes with extreme differential expression.

When D is small, a larger sample size is required to have power sufficient to detect these small differences in expression.

You can enter a range of values such as '1 2 3' or 0.2 to 2 by 0.1.

S1 (Standard Deviation Group 1)

Specify the standard deviation of expression in group 1. This standard deviation is assumed for all tests.

S1 and S2 should be on the the same scale as D.

To obtain the standard deviation of expression from the standard deviation of paired differences, use SDexpression = SDpaired/(sqrt(2)).

Because the true variation in paired differences will vary from gene to gene, it is recommended that a range of values be entered here.

You can enter a range of values such as *1 2 3* or *1 to 10 by 1*.

S2 (Standard Deviation Group 2)

Enter an estimate of the standard deviation of group 2. The standard deviation must be a positive number.

Enter S1 if you want to use the same value(s) as those for group 1.

Press the Standard Deviation Estimator button to obtain help on estimating the standard deviation.

You can enter a range of values such as *1 2 3* or *1 to 10 by 1*.

Number of Genes

Number of Genes Tested

Specify the number of genes for which hypothesis tests will be made.

This number will usually be the number of genes summarized on each array minus the number of housekeeping genes.

Only one number may be entered in this box.

Number of Genes – FDR Only

K (Number of Genes with Mean Difference > D)

Specify the number of genes for which a true mean difference in expression greater than D is expected.

K for EWER

The choice of K has no direct effect on the calculation of power or sample size when the False Discovery (Alpha) Method is set to EWER. K is not used when False Discovery (Alpha) method is set to EWER.

K for FDR

The choice of K has direct effect on the calculation of power or sample size when the False Discovery (Alpha) Method is set to FDR.

You can enter a range of values such as *10 20 30 40 50* or *20 to 100 by 10*.

Test

Alternative

Specify whether the hypothesis test for each gene is one-sided (directional) or two-sided (non-directional).

Recommendation: In most two-group experiments, differential expression in either direction (up-regulation or down-regulation) is of interest. Such experiments should have the Two-Sided alternative hypothesis.

For experiments for determining only whether expression has increased (or only decreased), a One-Sided alternative hypothesis is recommended. Often regulations dictate that the FDR or EWER level be divided by 2 for One-Sided alternative tests.

Test Type

Select the Test Statistic that will be used when the analysis of the gene expression data is carried out.

- **T**

  The T-Test assumes the expression values come from a normal distribution with UNKNOWN standard deviation (i.e., a standard deviation that will be estimated from the data).

- **Z**

  The Z-Test assumes the expression values come from a normal distribution with KNOWN standard deviation.

Recommendation: Because it very rare to know the true standard deviation of expression values in advance, T is the recommended test statistic.

Nonparametric Adjustment

Specify whether to make an adjustment for the Wilcoxon or Mann-Whitney (nonparametric) test.

The size of the adjustment depends on the assumed distribution. Select a distribution similar in shape to that of your data.

Options Tab

The Options tab contains convergence and iteration options that are rarely changed.

Maximum Iterations

Maximum Iterations Before Search Termination

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is left blank.

RECOMMENDED: 500 (or more)

Convergence Options

FDR Power Convergence

When FDR is selected for False Discovery (Alpha) Method, and Find (Solve For) is set to Power, the corresponding search algorithm will converge when the search criteria is below this value.

This value will rarely be changed from the default value.

RECOMMENDED: 0.0000000001

Example 1 – Finding Power

This example examines the power to detect differential expression for an experiment comparing a treatment group to a control group. There were 16 arrays used in each group. Each microarray produced intensity information for 5,000 genes. The 32 arrays were pre-processed by converting each expression value to the Log2 scale. In this example, the 2 Group T-Test was used to determine which genes were differentially expressed (upward or downward) when comparing the treatment group to the control group.

The researchers found very few differentially expressed genes, and wish to examine the power of the experiment to detect two-fold differential expression (Log2-scale difference of 1). Typical standard deviations in each group ranged from 0.2 to 2.0.

The researchers guess the number of genes with at least 2-fold differential expression to be around 50, but will examine the effect of this estimate on power by trying 10 and 100 genes as well. A false discovery rate of 0.05 was used.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Two-Sample T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **Two-Sample Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example1** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Power and Beta**
Power	*Ignored since this is the Find setting*
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.05**
N1 (Number of Arrays, Group 1)	**16**
N2 (Number of Arrays, Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
D (Difference)	**1.0**
S1 (Standard Deviation Group 1)	**0.2 to 2 by .2**
S2 (Standard Deviation Group 2)	**S1**
Number of Genes Tested	**5000**
K	**10 50 100**
Alternative Hypothesis	**Two-Sided**
Test Statistic	**T**
Nonparametric Adjustment	**Ignore**
Reports Tab	
Numeric Reports	**All Checked**
Number of Summary Statements	**1**
Show Plots	**Checked**
Interactive Format	**Unchecked**
Show Beta as Power	**Checked**

Annotated Output

Click the Run button to perform the calculations and generate the following output. The calculations should take a few moments.

Numeric Results

```
Numeric Results for Microarray Two-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
The standard deviations were assumed to be unknown and equal.
Number of Genes Tested: 5000

                                                             Single     Prob To
                                                               Gene      Detect
Power     N1/N2       D      S1      S2       K      FDR      Alpha       All K      Beta
1.00000   16/16     1.0     0.2     0.2      10   0.0500  0.0001055     1.00000   0.00000
1.00000   16/16     1.0     0.2     0.2      50   0.0500  0.0005316     1.00000   0.00000
1.00000   16/16     1.0     0.2     0.2     100   0.0500  0.0010741     1.00000   0.00000
0.98866   16/16     1.0     0.4     0.4      10   0.0500  0.0001043     0.89217   0.01134
0.99795   16/16     1.0     0.4     0.4      50   0.0500  0.0005305     0.90250   0.00205
0.99916   16/16     1.0     0.4     0.4     100   0.0500  0.0010732     0.91949   0.00084
0.52073   16/16     1.0     0.6     0.6      10   0.0500  0.0000549     0.00147   0.47927
0.75206   16/16     1.0     0.6     0.6      50   0.0500  0.0003998     0.00000   0.24794
0.83005   16/16     1.0     0.6     0.6     100   0.0500  0.0008916     0.00000   0.16995
0.06242   16/16     1.0     0.8     0.8      10   0.0500  0.0000066     0.00000   0.93758
0.23537   16/16     1.0     0.8     0.8      50   0.0500  0.0001251     0.00000   0.76463
0.34928   16/16     1.0     0.8     0.8     100   0.0500  0.0003752     0.00000   0.65072
.             .       .       .       .       .        .          .           .         .
.             .       .       .       .       .        .          .           .         .
.             .       .       .       .       .        .          .           .         .
```

References

Jung, S.-H. 2005. Sample size for FDR-control in microarray data analysis. Bioinformatics: Vol. 21 no. 14, pp. 3097-3104. Oxford University Press.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, MA.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions

Power is the individual probability of detecting each gene with true mean difference > D.
N1 and N2 are the number of arrays in groups 1 and 2 required to achieve the corresponding power.
D is the smallest difference in expression for which this power and sample size are valid.
S1 and S2 are the standard deviations for groups 1 and 2 used in each test.
K is the number of genes with true mean difference > D.
FDR is the expected proportion of false declarations of differential expression to total declarations of differential expression.
Single Gene Alpha is the probability of falsely declaring differential expression for an individual gene.
Prob to Detect All K is the probability of declaring differential expression for all K genes that have true mean difference > D.
Beta is the individual probability of failing to detect each gene with true mean difference > D.

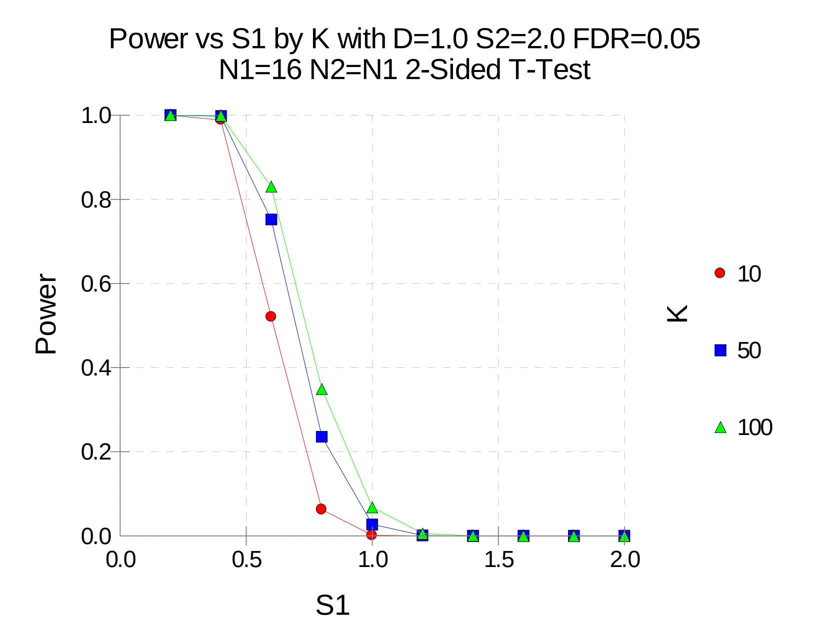
Summary Statements

Group sample sizes of 16 and 16 achieve 100% power for each gene to detect a true difference in expression of at least 1.0 with estimated group standard deviations of 0.2 and 0.2 and with a false discovery rate (FDR) of 0.0500 using a two-sided two-sample T-Test. For a single test, the individual test alpha is 0.0001055. The probability of detecting all 10 genes with true mean difference in expression > 1.0, is 1.00000.

This report shows the values of each of the parameters, one scenario per row. The values of power and beta were calculated from the other parameters.

The definitions of each column are given in the Report Definitions section.

Plots Section



This plot shows the relationship between power and the standard deviation of the differences for various three values of K. When the standard deviation within each group is greater than 1.0, the tests have very little power to detect 2-fold differences.

Example 2 – Finding the Sample Size

This example determines the number of arrays needed to achieve 80% power to detect differential expression for each gene. Each microarray will produce intensity information for 22,452 genes. The arrays will be pre-processed by converting each expression value to the Log2 scale. The two-sample T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment.

The researchers wish to detect differential expression that is two-fold or greater (Log2-scale difference of 1). Typical standard deviations in each group are expected to range from 0.2 to 2.0.

The researchers guess the number of genes with at least 2-fold differential expression to be around 50, but will examine the effect of this estimate on sample size by trying 10 and 100 genes as well. A false discovery rate of 0.05 will be used.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Two-Sample T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **Two-Sample Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example2** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1 (Group 1 Sample Size)**
Power	**0.80**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.05**
N1 (Number of Arrays, Group 1)	*Ignored since this is the Find setting*
N2 (Number of Arrays, Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
D (Difference)	**1.0**
S1 (Standard Deviation Group 1)	**0.2 to 2 by .2**
S2 (Standard Deviation Group 2)	**S1**
Number of Genes Tested	**22452**
K	**10 50 100**
Alternative Hypothesis	**Two-Sided**
Test Statistic	**T**
Nonparametric Adjustment	**Ignore**
Reports Tab	
Numeric Reports	**All Checked**
Number of Summary Statements	**1**
Show Plots	**Checked**
Interactive Format	**Unchecked**
Show Beta as Power	**Checked**

Output

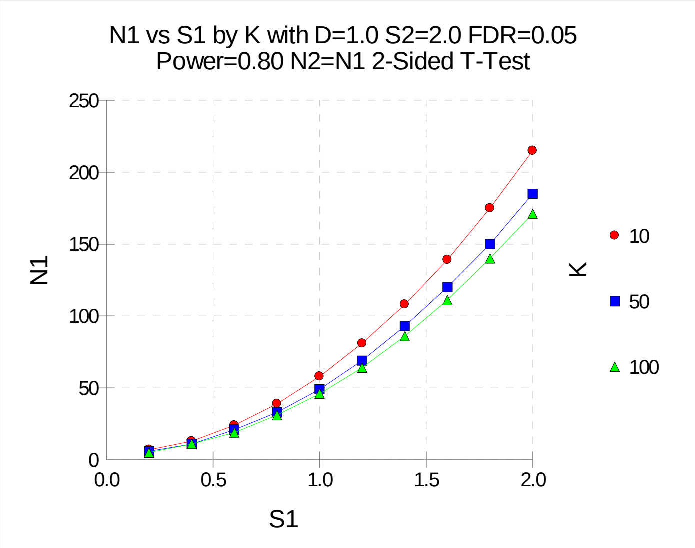
Click the Run button to perform the calculations and generate the following output. The calculations may take a few moments.

Numeric Results

```
Numeric Results for Microarray Two-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
The standard deviations were assumed to be unknown and equal.
Number of Genes Tested: 22452

                                                          Single   Prob To
                                                            Gene    Detect
Power     N1/N2       D     S1      S2      K      FDR     Alpha     All K      Beta
0.93967     7/7     1.0    0.2     0.2     10   0.0500 0.0000188   0.53673   0.06033
0.92971     6/6     1.0    0.2     0.2     50   0.0500 0.0000940   0.02615   0.07029
0.80449     5/5     1.0    0.2     0.2    100   0.0500 0.0001884   0.00000   0.19551
0.81237   13/13     1.0    0.4     0.4     10   0.0500 0.0000188   0.12518   0.18763
0.80047   11/11     1.0    0.4     0.4     50   0.0500 0.0000940   0.00001   0.19953
0.86440   11/11     1.0    0.4     0.4    100   0.0500 0.0001884   0.00000   0.13560
0.82116   24/24     1.0    0.6     0.6     10   0.0500 0.0000188   0.13940   0.17884
0.83607   21/21     1.0    0.6     0.6     50   0.0500 0.0000940   0.00013   0.16393
0.81695   19/19     1.0    0.6     0.6    100   0.0500 0.0001884   0.00000   0.18305
0.81806   39/39     1.0    0.8     0.8     10   0.0500 0.0000188   0.13424   0.18194
0.80753   33/33     1.0    0.8     0.8     50   0.0500 0.0000940   0.00002   0.19247
0.81606   31/31     1.0    0.8     0.8    100   0.0500 0.0001884   0.00000   0.18394
.             .       .      .       .      .        .         .         .         .
.             .       .      .       .      .        .         .         .         .
.             .       .      .       .      .        .         .         .         .
```

This report shows the values of each of the parameters, one scenario per row. The sample size (number of arrays) estimates were calculated from the other parameters. The power is the actual power produced by the given sample size.

Plots Section



This plot shows the relationship between sample size and the standard deviations within each group for three values of K.

Example 3 – Finding the Minimum Detectable Difference

This example finds the minimum difference in expression that can be detected with 90% power from a microarray experiment with two groups of 9 arrays in each group. The 9 arrays permit tests on 7,228 genes. The arrays will be pre-processed by converting each expression value to the Log2 scale. The two-sample T-test will be used to determine which genes are differentially expressed (upward or downward) following exposure to the treatment. Typical standard deviations in each group for this experiment range from 0.2 to 1.8.

In this example we will examine a range for K (the number of genes with mean difference greater than the minimum detectable difference), since this should vary with the mean difference chosen. A false discovery rate of 0.05 will be used.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Two-Sample T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **Two-Sample Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example3** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**Mean Difference**
Power	**0.90**
False Discovery (Alpha) Method	**FDR (False Discovery Rate)**
FDR or EWER Value	**0.05**
N1 (Number of Arrays, Group 1)	**9**
N2 (Number of Arrays, Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
D (Difference)	*Ignored since this is the Find setting*
S1 (Standard Deviation Group 1)	**0.2 to 1.8 by .4**
S2 (Standard Deviation Group 2)	**S1**
Number of Genes Tested	**7228**
K	**10 to 50 by 10**
Alternative Hypothesis	**Two-Sided**
Test Statistic	**T**
Nonparametric Adjustment	**Ignore**
Reports Tab	
Numeric Reports	**All Checked**
Number of Summary Statements	**1**
Show Plots	**Checked**
Interactive Format	**Unchecked**
Show Beta as Power	**Checked**
Mean Difference and SD Decimals	**4**

Output

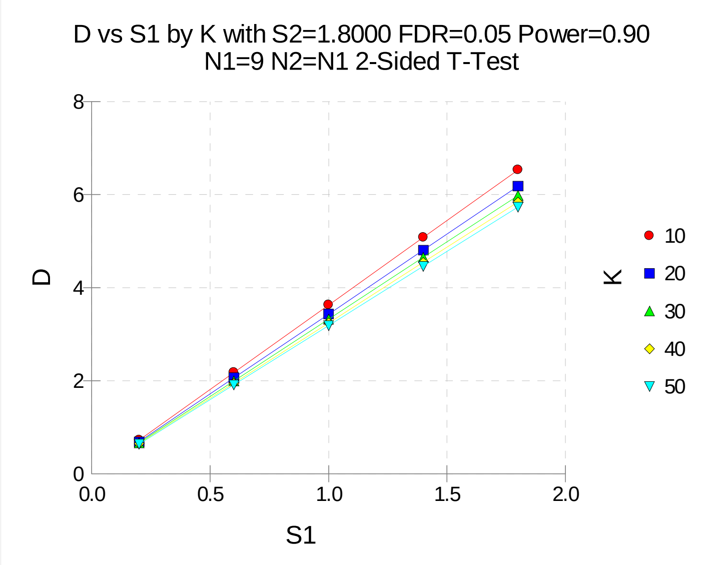
Click the Run button to perform the calculations and generate the following output. The calculations may take a few moments.

Numeric Results

```
Numeric Results for Microarray Two-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
The standard deviations were assumed to be unknown and equal.
Number of Genes Tested: 7228

                                                                 Single    Prob To
                                                                   Gene     Detect
Power    N1/N2       D      S1       S2       K       FDR         Alpha      All K        Beta
0.90000    9/9  0.6626  0.2000   0.2000      10    0.0500     0.0000656    0.34868     0.10000
0.90000    9/9  0.6253  0.2000   0.2000      20    0.0500     0.0001314    0.12158     0.10000
0.90000    9/9  0.6038  0.2000   0.2000      30    0.0500     0.0001974    0.04239     0.10000
0.90000    9/9  0.5888  0.2000   0.2000      40    0.0500     0.0002636    0.01478     0.10000
0.90000    9/9  0.5772  0.2000   0.2000      50    0.0500     0.0003300    0.00515     0.10000
0.90000    9/9  1.9879  0.6000   0.6000      10    0.0500     0.0000656    0.34868     0.10000
0.90000    9/9  1.8759  0.6000   0.6000      20    0.0500     0.0001314    0.12158     0.10000
0.90000    9/9  1.8115  0.6000   0.6000      30    0.0500     0.0001974    0.04239     0.10000
0.90000    9/9  1.7663  0.6000   0.6000      40    0.0500     0.0002636    0.01478     0.10000
0.90000    9/9  1.7315  0.6000   0.6000      50    0.0500     0.0003300    0.00515     0.10000
0.90000    9/9  3.3132  1.0000   1.0000      10    0.0500     0.0000656    0.34868     0.10000
0.90000    9/9  3.1265  1.0000   1.0000      20    0.0500     0.0001314    0.12158     0.10000
0.90000    9/9  3.0192  1.0000   1.0000      30    0.0500     0.0001974    0.04239     0.10000
.            .       .       .        .       .         .             .          .           .
.            .       .       .        .       .         .             .          .           .
.            .       .       .        .       .         .             .          .           .
```

This report shows the values of each of the parameters, one scenario per row. The Minimum Mean Difference (D) estimates were calculated from the other parameters.

Plots Section



This plot shows the relationship between D (the minimum detectable difference on the Log2 scale) and the standard deviations within each group for five values of K.

Example 4 – Validation (EWER) using Stekel

Stekel (2003), page 228, gives an example in which Power = 0.95, D = 1, and S1 = S2 = 0.68 for a two-sided two-sample T-Test. The number of genes tested is 10000. The control of false discoveries is "at most one false positive result." This corresponds to an EWER value of 1.0. The sample sizes obtained for this example are 33 per group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Two-Sample T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **Two-Sample Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example4** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1 (Group 1 Sample Size)**
Power	**0.95**
False Discovery (Alpha) Method	**EWER (Experiment-Wise Error Rate)**
FDR or EWER Value	**1.0**
N1 (Number of Arrays, Group 1)	*Ignored since this is the Find setting*
N2 (Number of Arrays, Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
D (Difference)	**1.0**
S1 (Standard Deviation Group 1)	**0.68**
S2 (Standard Deviation Group 2)	**S1**
Number of Genes Tested	**10000**
K	*Ignored since EWER is used*
Alternative Hypothesis	**Two-Sided**
Test Statistic	**T**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Microarray Two-Sample T-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
The standard deviations were assumed to be unknown and equal.
Number of Genes Tested: 10000
                                                                      Single
                                                                        Gene
Power           N1/N2           D         S1         S2        EWER    Alpha          Beta
0.95785         33/33        1.00       0.68       0.68      1.0000    0.0001000    0.04215
```

The sample sizes of 33 per group match Stekel's result.

Example 5 – Validation (EWER) using Lee

Lee (2004), pp. 218-220, gives an example in which Power = 0.90, D = 1.0 1.5 2.0 2.5 and S = 1.0 for a two-sided Z-Test. The corresponding S1 and S2 for a two-sample design is $1.0/\sqrt{2}$ = 0.707107. The number of genes tested is 1000. The control of false discoveries is 0.5. This corresponds to an EWER value of 0.5. This setup corresponds to the upper left corner of Table 14.3 on page 219. The sample sizes obtained for this setup are 23, 11, 6, and 4, respectively.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Two-Sample T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **Two-Sample Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example5** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1 (Group 1 Sample Size)**
Power	**0.90**
False Discovery (Alpha) Method	**EWER (Experiment-Wise Error Rate)**
FDR or EWER Value	**0.5**
N1 (Number of Arrays, Group 1)	*Ignored since this is the Find setting*
N2 (Number of Arrays, Group 2)	**Use R**
R (Sample Allocation Ratio)	**1.0**
D (Difference)	**1.0 1.5 2.0 2.5**
S1 (Standard Deviation Group 1)	**0.707107**
S2 (Standard Deviation Group 2)	**S1**
Number of Genes Tested	**1000**
K	*Ignored since EWER is used*
Alternative Hypothesis	**Two-Sided**
Test Statistic	**Z**
Nonparametric Adjustment	**Ignore**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Microarray Two-Sample Z-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff <> 0
The standard deviations were assumed to be known and equal.
Number of Genes Tested: 1000

                                                                  Single
                                                                    Gene
Power          N1/N2          D         S1         S2     EWER     Alpha        Beta
0.90576        23/23   1.000000   0.707107   0.707107   0.5000 0.0005000     0.09424
0.93244        11/11   1.500000   0.707107   0.707107   0.5000 0.0005000     0.06756
0.92194          6/6   2.000000   0.707107   0.707107   0.5000 0.0005000     0.07806
0.93565          4/4   2.500000   0.707107   0.707107   0.5000 0.0005000     0.06435
```

Sample sizes of 23, 11, 6, and 4 per group match the results shown in Lee (2004).

Example 6 – Validation (FDR) using Jung

Jung (2005), page 3100, gives an example for the sample size needed to control FDR in a two-sample Z-Test. This example is repeated in Chow, Shao, and Wang (2008). In the example, Power = 0.60 (from 24/40), D = 1.0, and S = 1.0 for a one-sided two-sample Z-Test. The number of genes tested is 4000. The FDR level is 1%. This setup corresponds to Example 1 on page 3100. The required sample size obtained in each group for this setup is 34.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Two-Sample T-Test for Microarray Data** procedure window by clicking on **Microarray**, then **Two-Sample Tests**. You may then follow along here by making the appropriate entries as listed below or load the completed template **Example6** from the Template tab on the procedure window.

Option	Value
Data Tab	
Find (Solve For)	**N1 (Group 1 Sample Size)**
Power	**0.60**
False Discovery (Alpha) Method.............	**FDR (False Discovery Rate)**
FDR or EWER Value...............................	**0.01**
N (Number of Arrays)	*Ignored since this is the Find setting*
Alternative Hypothesis	**One-Sided**
Test Statistic...	**Z**
D (Difference)..	**1.0**
S1 (Standard Deviation Group 1)............	**1.0**
S2 (Standard Deviation Group 2)............	**S1**
Number of Genes Tested........................	**4000**
K	**40**

Output

Click the Run button to perform the calculations and generate the following output.

Numeric Results

```
Numeric Results for Microarray Two-Sample Z-Test
Null Hypothesis: MeanDiff = 0     Alternative Hypothesis: MeanDiff > 0
The standard deviations were assumed to be known and equal.
Number of Genes Tested: 4000

                                                             Single    Prob To
                                                               Gene     Detect
Power    N1/N2        D       S1         S2        K       FDR     Alpha      All K        Beta
0.61099  34/34   1.0000  1.0000     1.0000       40    0.0100 0.0000612    0.00000     0.38901
```

A group sample size of 34 matches the result shown in Jung (2005). For Example 3 in Jung (2005), the alternative hypothesis is two-sided and results in a sample size of 73. This result may be validated in ***PASS*** by changing Alternative to Two-Sided in this example.

References

A

A’Hern, R. P. A. 2001. "Sample size tables for exact single-stage phase II designs." *Statistics in Medicine*, Volume 20, pages 859-866.

Al-Sunduqchi, Mahdi S. 1990. *Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics*. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

Armitage, P., and Colton, T. 1998. *Encyclopedia of Biostatistics*. John Wiley, New York.

Armitage, P., McPherson, C.K., and Rowe, B.C. 1969. “Repeated significance tests on accumulating data.” *Journal of the Royal Statistical Society, Series A*, 132, pages 235-244.

Atkinson, A.C. 1985. *Plots, Transformations, and Regression*. Oxford University Press, Oxford (also in New York). This book goes into the details of regression diagnostics and plotting. It puts together much of the recent work in this area.

Atkinson, A.C., and Donev, A.N. 1992. *Optimum Experimental Designs*. Oxford University Press, Oxford. This book discusses D-Optimal designs.

B

Bain, L.J. and Engelhardt, M. 1991. *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker. New York. This book contains details for testing data that follow the exponential and Weibull distributions.

Barnard, G.A. 1947. "Significance tests for 2 x 2 tables." *Biometrika* 34:123-138.

Bartholomew, D.J. 1963. "The Sampling Distribution of an Estimate Arising in Life Testing." *Technometrics*, Volume 5 No. 3, 361-374.

Beal, S. L. 1987. "Asymptotic Confidence Intervals for the Difference between Two Binomial Parameters for Use with Small Samples." *Biometrics*, Volume 43, Issue 4, 941-950.

Benjamini, Y. and Hochberg, Y. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society, Series B (Methodological),* Vol. 57, No. 1, 289-300.

Blackwelder, W.C. 1993. "Sample size and power in prospective analysis of relative risk." *Statistics in Medicine*, Volume 12, 691-698.

Blackwelder, W.C. 1998. "Equivalence Trials." In *Encyclopedia of Biostatistics*, John Wiley and Sons. New York. Volume 2, 1367-1372.

Bonett, Douglas. 2002. “Sample Size Requirements for Testing and Estimating Coefficient Alpha.” *Journal of Educational and Behavioral Statistics*, Vol. 27, pages 335-340.

Box, G.E.P., Hunter, S. and Hunter, J.S.. 1978. *Statistics for Experimenters*. John Wiley & Sons, New York. This is probably the leading book in the area experimental design in industrial experiments. You definitely should acquire and study this book if you plan anything but a casual acquaintance with experimental design. The book is loaded with examples and explanations.

Breslow, N. E. and **Day, N. E.** 1980. *Statistical Methods in Cancer Research: Volume 1. The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Brown, H., and Prescott, R. 2006. *Applied Mixed Models in Medicine*. 2nd ed. John Wiley & Sons Ltd. Chichester, West Sussex, England.

Brush, Gary G. 1988. *Volume 12: How to Choose the Proper Sample Size*, American Society for Quality Control, 310 West Wisconsin Ave, Milwaukee, Wisconsin, 53203. This is a small workbook for quality control workers.

C

Casagrande, J. T., Pike, M.C., and Smith, P. G. 1978. “The Power Function of the “Exact” Test for Comparing Two Binomial Distributions,” *Applied Statistics*, Volume 27, No. 2, pages 176-180. This article presents the algorithm upon which our Fisher’s exact test is based.

Cochran and Cox. 1992. *Experimental Designs. Second Edition*. John Wiley & Sons. New York. This is one of the classic books on experimental design, first published in 1957.

Chen, K.W.; Chow, S.C.; and Li, G. 1997. "A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs" *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 25, No. 6, pages 753-765.

Chen, T. T. 1997. "Optimal Three-Stage Designs for Phase II Cancer Clinical Trials." *Statistics in Medicine*, Volume 16, pages 2701-2711.

Chow, S.C. and Liu, J.P. 1999. *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker. New York.

Chow, S.C.; Shao, J.; Wang, H. 2003. *Sample Size Calculations in Clinical Research*. Marcel Dekker. New York.

Chow, S.-C.; Shao, J.; Wang, H. 2008. *Sample Size Calculations in Clinical Research, Second Edition*. Chapman & Hall/CRC. Boca Raton, Florida.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey. This is a very nice, clearly written book. There are MANY examples. It is the largest of the sample size books. It does not deal with clinical trials.

Cohen, Jacob. 1990. “Things I Have Learned So Far.” *American Psychologist*, December, 1990, pages 1304-1312. This is must reading for anyone still skeptical about the need for power analysis.

Collett, D. 1991. *Modelling Binary Data.* Chapman & Hall, New York, New York. This book covers such topics as logistic regression, tests of proportions, matched case-control studies, and so on.

Collett, D. 1994. *Modelling Survival Data in Medical Research.* Chapman & Hall, New York, New York. This book covers such survival analysis topics as Cox regression and log rank tests.

Conlon, M. and Thomas, R. 1993. “The Power Function for Fisher’s Exact Test.” *Applied Statistics*, Volume 42, No. 1, pages 258-260. This article was used to validate the power calculations of Fisher’s Exact Test in PASS. Unfortunately, we could not use the algorithm to improve the speed because the algorithm requires equal sample sizes.

Cook, R. D. and **Weisburg, S.** 1999. Applied Regression Including Computing and Graphics. John Wiley and Sons, Inc.

Cox, D. R. 1972. “Regression Models and life tables.” *Journal of the Royal Statistical Society, Series B*, Volume 34, Pages 187-220. This article presents the proportional hazards regression model.

D

D'Agostino, R.B., Chase, W., Belanger, A. 1988."The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations.", *The American Statistician*, August 1988, Volume 42 Number 3, pages 198-202.

Davies, Owen L. 1971. *The Design and Analysis of Industrial Experiments*. Hafner Publishing Company, New York. This was one of the first books on experimental design and analysis. It has many examples and is highly recommended.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. 1988. “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics*, 44, pages 837-845.

DeMets, D.L. and Lan, K.K.G. 1984. “An overview of sequential methods and their applications in clinical trials.” *Communications in Statistics, Theory and Methods*, 13, pages 2315-2338.

DeMets, D.L. and Lan, K.K.G. 1994. “Interim analysis: The alpha spending function approach.” *Statistics in Medicine*, 13, pages 1341-1352.

Demidenko, E. 2004. *Mixed Models – Theory and Applications*. John Wiley & Sons. Hoboken, New Jersey.

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. Academic Press. New York. (Presents many useful results for determining sample sizes.)

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation.* Springer-Verlag. New York. This book is currently available online at http://jeff.cs.mcgill.ca/~luc/rnbookindex.html.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. 1994. *Analysis of Longitudinal Data*. Oxford University Press. New York, New York.

Dixon, W. J. and Tukey, J. W. 1968. “Approximate behavior of the distribution of Winsorized t,” *Technometrics*, Volume 10, pages 83-98.

Donnelly, Thomas G. 1980. “ACM Algorithm 462: Bivariate Normal Distribution,” *Collected Algorithms from ACM*, Volume II, New York, New York.

Donner, Allan. 1984. “Approaches to Sample Size Estimation in the Design of Clinical Trials--A Review,” *Statistics in Medicine*, Volume 3, pages 199-214. This is a well done review of the clinical trial literature. Although it is becoming out of date, it is still a good place to start.

Donner, A. and Klar, N. 1996. “Statistical Considerations in the Design and Analysis of Community Intervention Trials.” *The Journal of Clinical Epidemiology*, Vol. 49, No. 4, 1996, pages 435-439.

Donner, A. and Klar, N. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold. London.

Draghici, S. 2003. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC. London. This is an excellent overview of most areas of Microarray analysis.

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. 2002. “Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Experiments,” *Statistica Sinica*, Volume 12, pages 111-139.

Dunn, O. J. 1961. “Multiple comparisons among means,” *Journal of the American Statistical Association*, Volume 56, pages 52-64.

Dunn, O. J. 1964. “Multiple comparisons using rank sums,” *Technometrics*, Volume 6, pages 241-252.

Dunnett, C. W. 1955. “A Multiple comparison procedure for Comparing Several Treatments with a Control,” *Journal of the American Statistical Association*, Volume 50, pages 1096-1121.

Dupont, William. 1988. “Power Calculations for Matched Case-Control Studies,” *Biometrics*, Volume 44, pages 1157-1168.

Dupont, William and Plummer, Walton D. 1990. “Power and Sample Size Calculations--A Review and Computer Program,” *Controlled Clinical Trials*, Volume 11, pages 116-128. Documents a nice public-domain program on sample size and power analysis.

E

Edgington, E. 1987. *Randomization Tests*. Marcel Dekker. New York. A comprehensive discussion of randomization tests with many examples.

Edwards, L.K. 1993. *Applied Analysis of Variance in the Behavior Sciences*. Marcel Dekker. New York. Chapter 8 of this book is used to validate the repeated measures module of PASS.

Efron, B. 1971. “Forcing a Sequential Experiment to be Balanced.” *Biometrika*. Volume 58, pages 403-417.

Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall. New York.

Elandt-Johnson, R.C. and Johnson, N.L. 1980. *Survival Models and Data Analysis*. John Wiley. NY, NY. This book devotes several chapters to population and clinical life-table analysis.

Epstein, Benjamin. 1960. “Statistical Life Test Acceptance Procedures.” *Technometrics*. Volume 2.4, pages 435-446.

F

Farrington, C. P. and Manning, G. 1990. “Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.” *Statistics in Medicine*, Vol. 9, pages 1447-1454. This article contains the formulas used for the Equivalence of Proportions module in PASS.

Feldt, L.S., Woodruff, D.J., and Salih, F.A. 1987. “Statistical inference for coefficient alpha.” *Applied Psychological Measurement*, Vol. 11, pages 93-103.

Feldt, L.S. and Ankenmann, R.D. 1999. “Determining Sample Size for a Test of the Equality of Alpha Coefficients When the Number of Part-Tests is Small.” *Psychological Methods*, Vol. 4(4), pages 366-377.

Fisher, R. A. 1921. “On the probable error of a coefficient of correlation deduced from a small sample.” *Metron*, i (4), 1-32.

Flack, V. F., Afifi, A. A., Lachenbruch, P. A., and Schouten, H. J. A. 1988. "Sample Size Determinations for the Two Rater Kappa Statistic." *Psychometrika*, Volume 53, No. 3, pages 321-325.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons. New York. This book provides a very good introduction to clinical trials. It may be a bit out of date now, but it is still very useful.

Fleiss, J. L., Levin, B., Paik, M.C. 2003. *Statistical Methods for Rates and Proportions. Third Edition*. John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleming, T. R. 1982. "One-sample multiple testing procedure for Phase II clinical trials." *Biometrics*, Volume 38, pages 143-151.

Freedman, L.S. 1982. "Tables of the Number of Patients Required in Clinical Trials using the Logrank Test." *Statistics in Medicine*, 1:121-129.

G

Gans. 1984. “The Search for Significance: Different Tests on the Same Data.” *The Journal of Statistical Computation and Simulation*, 1984, pages 1-21.

Gart, John J. and Nam, Jun-mo. 1988. "Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness." *Biometrics*, Volume 44, Issue 2, 323-338.

Gart, John J. and Nam, Jun-mo. 1990. "Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables." *Biometrics*, Volume 46, Issue 3, 637-643.

Gehlback, Stephen. 1988. *Interpreting the Medical Literature: Practical Epidemiology for Clinicians*. Second Edition. McGraw-Hill. New York. Telephone: (800)722-4726. The preface of this book states that its purpose is to provide the reader with a useful approach to interpreting the quantitative results given in medical literature. We reference it specifically because of its discussion of ROC curves.

Gentle, James E. 1998. *Random Number Generation and Monte Carlo Methods*. Springer. New York.

Gibbons, J. 1976. *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart and Winston. New York.

Gibbons, J. 1985. *Nonparametric Methods for Quantitative Analysis (2nd Edition)*. American Sciences Press. New York.

Goldstein, Richard. 1989. “Power and Sample Size via MS/PC-DOS Computers,” *The American Statistician*, Volume 43, Number 4, pages 253-260. A comparative review of power analysis software that was available at that time.

Greenwood, J. A. and **Sandomire, M. M.** 1950. “Sample Size Required for Estimating the Standard Deviation as a Per Cent of its True Value”, *Journal of the American Statistical Association*, Vol. 45, No. 250, pp. 257-260.

Griffiths, P. and Hill, I.D. 1985. *Applied Statistics Algorithms*, The Royal Statistical Society, London, England. See page 243 for ACM algorithm 291.

Gross and Clark 1975. *Survival Distributions*: Reliability Applications in Biomedical Sciences. John Wiley, New York.

Guenther, William C. 1977. “Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient,” *The American Statistician*, Volume 31, Number 1, pages 45-48.

Guenther, William C. 1977. *Sampling Inspection in Statistical Quality Control*. Griffin’s Statistical Monographs, Number 37. London.

H

Hahn, G. J. and Meeker, W.Q. 1991. *Statistical Intervals*. John Wiley & Sons. New York.

Hanley, J. A. and McNeil, B. J. 1982. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, 143, 29-36. April, 1982.

Hanley, J. A. and McNeil, B. J. 1983. "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases." *Radiology*, 148, 839-843. September, 1983.

Harris, M., Horvitz, D. J., and **Mood, A. M.** 1948. “On the Determination of Sample Sizes in Designing Experiments”, Journal of the American Statistical Association, Volume 43, No. 243, pp. 391-402.

Hernandez-Bermejo, B. and Sorribas, A. 2001. “Analytical Quantile Solution for the S-distribution, Random Number Generation and Statistical Data Modeling.” *Biometrical Journal* 43, 1007-1025.

Hoaglin, Mosteller, and Tukey. 1985. *Exploring Data Tables, Trends, and Shapes*. John Wiley. New York.

Hochberg, Y. and Tamhane, A. C. 1987. *Multiple Comparison Procedures*. John Wiley & Sons. New York.

Howe, W.G. 1969. “Two-Sided Tolerance Limits for Normal Populations—Some Improvements.” *Journal of the American Statistical Association*, 64, 610-620.

Hosmer, D. and Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons. New York. This book gives an advanced, in depth look at logistic regression.

Hosmer, D. and Lemeshow, S. 1999. *Applied Survival Analysis*. John Wiley & Sons. New York.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24, 417-441, 498-520.

Hsieh, F.Y. 1989. “Sample Size Tables for Logistic Regression,” *Statistics in Medicine*, Volume 8, pages 795-802. This is the article that was the basis for the sample size calculations in logistic regression in PASS 6.0. It has been superceded by the 1998 article.

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. “A Simple Method of Sample Size Calculation for Linear and Logistic Regression,” *Statistics in Medicine*, Volume 17, pages 1623-1634. The sample size calculation for logistic regression in PASS are based on this article.

Hsieh, F.Y. and Lavori, P.W. 2000. “Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates,” *Controlled Clinical Trials*, Volume 21, pages 552-560. The sample size calculation for Cox regression in PASS are based on this article.

Hsu, Jason. 1996. *Multiple Comparisons: Theory and Methods.* Chapman & Hall. London. This book gives a beginning to intermediate discussion of multiple comparisons, stressing the interpretation of the various MC tests. It provides details of three popular MC situations: all pairs, versus the best, and versus a control. The power calculations used in the MC module of PASS came from this book.

J

Johnson, N.L., Kotz, S., and Kemp, A.W. 1992. *Univariate Discrete Distributions, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions Volume 1, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1995. *Continuous Univariate Distributions Volume 2, Second Edition*. John Wiley & Sons. New York.

Julious, Steven A. 2004. "Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data." *Statistics in Medicine*, 23:1921-1986.

Jung, Sin-Ho. 2005. "Sample size for FDR-control in microarray data analysis" *Bioinformatics*, 21(14):3097-3104.

Jung, Sin-Ho; Kang, Sun J.; McCall, Linda M.; Blumenstein, Brent. 2005**.** “Sample Sizes Computation for Two-Sample Noninferiority Log-Rank Test”, *J. of Biopharmaceutical Statistics,* Volume 15, pages 969-979.

Juran, J.M. 1979. *Quality Control Handbook*. McGraw-Hill. New York.

K

Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Karian, Z.A and Dudewicz, E.J. 2000. *Fitting Statistical Distributions.* CRC Press, New York.

Katz, D., Baptista, J., Azen, S. P., and Pike, M. C. 1978. “Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies,” *Biometrics*, 34, pages 469-474.

Kendall,M. and Stuart, A. 1987. *Kendall’s Advanced Theory of Statistics. Volume 1: Distribution Theory*. Oxford University Press. New York. This is a fine math-stat book for graduate students in statistics. We reference it because it includes formulas that are used in the program.

Kenward, M. G. and Roger, J. H. 1997. “Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood,” *Biometrics,* 53, pages 983-997.

Kirk, Roger E. 1982. *Experimental Design: Procedures for the Behavioral Sciences.* Brooks/Cole. Pacific Grove, California. This is a respected reference on experimental design and analysis of variance.

Klein, J.P. and Moeschberger, M.L.. 1997. *Survival Analysis.* Springer-Verlag. New York. This book provides a comprehensive look at the subject complete with formulas, examples, and lots of useful comments. It includes all the more recent developments in this field. I recommend it.

Koch, G.G.; Atkinson, S.S.; Stokes, M.E. 1986. *Encyclopedia of Statistical Sciences.* Volume 7. John Wiley. New York. Edited by Samuel Kotz and Norman Johnson. The article on Poisson Regression provides a very good summary of the subject.

Kraemer, H. C. and **Thiemann, S.** 1987. *How Many Subjects*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Kupper, L. L. and **Hafner, K. B.** 1989. 'How Appropriate are Popular Sample Size Formulas?', The American Statistician, Volume 43, No. 2, pp. 101-105.

L

Lachin, John M. 2000. *Biostatistical Methods.* John Wiley & Sons. New York. This is a graduate-level methods book that deals with statistical methods that are of interest to biostatisticians such as odds ratios, relative risks, regression analysis, case-control studies, and so on.

Lachin, John M. and **Foulkes, Mary A.** 1986**.** “Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification,” *Biometrics,* Volume 42, September, pages 507-516.

Lakatos, Edward. 1988**.** “Sample Sizes Based on the Log-Rank Statistic in Complex Clinical Trials”, *Biometrics,* Volume 44, March, pages 229-241.

Lakatos, Edward. 2002**.** “Designing Complex Group Sequential Survival Trials”, *Biometrika,* Volume 70, pages 1969-1989.

Lan, K.K.G. and DeMets, D.L. 1983. “Discrete sequential boundaries for clinical trials.” *Biometrika*, 70, pages 659-663.

Lan, K.K.G. and Zucker, D.M. 1993. “Sequential monitoring of clinical trials: the role of information and Brownian motion.” *Statistics in Medicine*, 12, pages 753-765.

Lance, G.N. and Williams, W.T. 1967. “A general theory of classificatory sorting strategies. I. Hierarchical systems.” *Comput. J.* 9, pages 373-380.

Lance, G.N. and Williams, W.T. 1967. “Mixed-data classificatory programs I. Agglomerative systems.” *Aust.Comput. J.* 1, pages 15-20.

Lawless, J.F. 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

Lawson, John. 1987. *Basic Industrial Experimental Design Strategies*. Center for Statistical Research at Brigham Young University. Provo, Utah. 84602. This is a manuscript used by Dr. Lawson in courses and workshops that he provides to industrial engineers. It is the basis for many of our experimental design procedures.

Lee, E.T. 1980. *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications. Belmont, California.

Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis*. Second Edition. John Wiley & Sons. New York. This book provides a very readable introduction to survival analysis techniques.

Lee, M.-L. T. 2004. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers. Norwell, Massachusetts.

Lenth, Russell V. 1987. “Algorithm AS 226: Computing Noncentral Beta Probabilities,” *Applied Statistics*, Volume 36, pages 241-244.

Lenth, Russell V. 1989. “Algorithm AS 243: Cumulative Distribution Function of the Non-central t Distribution,” *Applied Statistics*, Volume 38, pages 185-189.

Lewis, J.A. 1999. “Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline.” *Statistics in Medicine*, 18, pages 1903-1942.

Lipsey, Mark W. 1990. *Design Sensitivity Statistical Power for Experimental Research*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Littell, R. C. et al. 2006. *SAS for Mixed Models – Second Edition*. SAS Institute Inc., Cary, North Carolina.

Liu, J., Hsueh, H., Hsieh, E., and Chen, J.J. 2002. “Tests for equivalence or non-inferiority for paired binary data,” *Statistics in Medicine*, Volume 21, pages 231-245.

Liu, H. and Wu, T. 2005. "Sample Size Calculation and Power Analysis of Time-Averaged Difference," *Journal of Modern Applied Statistical Methods*, Vol. 4, No. 2, pages 434-445.

Lu, Y. and Bean, J.A. 1995. “On the sample size for one-sided equivalence of sensitivities based upon McNemar’s test,” *Statistics in Medicine*, Volume 14, pages 1831-1839.

Locke, C.S. 1984. “An exact confidence interval for untransformed data for the ratio of two formulation means,” *J. Pharmacokinet. Biopharm.*, Volume 12, pages 649-655.

M

Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. *Sample Size Tables for Clinical Studies, 2nd Edition*. Blackwell Science. Malden, Mass. A very good & easy to read book on determining appropriate sample sizes in many situations.

Marubini, E. and **Valsecchi, M.G.** 1996. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley: New York, New York.

Matsumoto, M. and Nishimura,T. 1998. “Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator” *ACM Trans. On Modeling and Computer Simulations*.

McClish, D.K. 1989. “Analyzing a Portion of the ROC Curve.” *Medical Decision Making*, 9: 190-195

Metz, C.E. 1978. “Basic principles of ROC analysis.” *Seminars in Nuclear Medicine*, Volume 8, No. 4, pages 283-298.

Miettinen, O.S. and Nurminen, M. 1985. “Comparative analysis of two rates.” *Statistics in Medicine* 4: 213-226.

Montgomery, Douglas. 1984. *Design and Analysis of Experiments*. John Wiley & Sons, New York. A textbook covering a broad range of experimental design methods. The book is not limited to industrial investigations, but gives a much more general overview of experimental design methodology.

Moore, D. S. and McCabe, G. P. 1999. *Introduction to the Practice of Statistics*. W. H. Freeman and Company. New York.

Moura, Eduardo C. 1991. *How To Determine Sample Size And Estimate Failure Rate in Life Testing*. ASQC Quality Press. Milwaukee, Wisconsin.

Mukerjee, H., Robertson, T., and Wright, F.T. 1987. “Comparison of Several Treatments With a Control Using Multiple Contrasts.” *Journal of the American Statistical Association*, Volume 82, No. 399, pages 902-910.

Muller, K. E., and Barton, C. N. 1989. “Approximate Power for Repeated-Measures ANOVA Lacking Sphericity.” *Journal of the American Statistical Association*, Volume 84, No. 406, pages 549-555.

Muller, K. E., LaVange, L.E., Ramey, S.L., and Ramey, C.T. 1992. “Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications.” *Journal of the American Statistical Association*, Volume 87, No. 420, pages 1209-1226.

Muller, K. E. and Stewart, P.W. 2006. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley & Sons Inc. Hoboken, New Jersey.

Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, Massachusetts. This is one of the bibles on the topic of regression analysis.

N

Nam, Jun-mo. 1987. “A Simple Approximation for Calculating Sample Sizes for Detecting Linear Trend in Proportions,” *Biometrics*, Volume 43, pages 701-705.

Nam, Jun-mo. 1992. “Sample Size Determination for Case-Control Studies and the Comparison of Stratified and Unstratified Analyses,” *Biometrics*, Volume 48, pages 389-395.

Nam, Jun-mo. 1997. “Establishing equivalence of two treatments and sample size requirements in matched-pairs design,” *Biometrics*, Volume 53, pages 1422-1430.

Nam, J-m. and Blackwelder, W.C. 2002. “Analysis of the ratio of marginal probabilities in a matched-pair setting,” *Statistics in Medicine*, Volume 21, pages 689-699.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. 1996. *Applied Linear Statistical Models*. Richard D. Irwin, Inc. Chicago, Illinois. This mammoth book covers regression analysis and analysis of variance thoroughly and in great detail. We recommend it.

Neter, J., Wasserman, W., and Kutner, M. 1983. *Applied Linear Regression Models*. Richard D. Irwin, Inc. Chicago, Illinois. This book provides you with a complete introduction to the methods of regression analysis. We suggest it to non-statisticians as a great reference tool.

Newcombe, Robert G. 1998a. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods." *Statistics in Medicine*, Volume 17, 857-872.

Newcombe, Robert G. 1998b. "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods." *Statistics in Medicine*, Volume 17, 873-890.

Newcombe, Robert G. 1998c. "Improved Confidence Intervals for the Difference Between Binomial Proportions Based on Paired Data." *Statistics in Medicine*, Volume 17, 2635-2650.

O

O'Brien, P.C. and Fleming, T.R. 1979. "A multiple testing procedure for clinical trials." *Biometrics*, 35, pages 549-556.

O'Brien, R.G. and Kaiser, M.K. 1985. "MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer." *Psychological Bulletin*, 97, pages 316-333.

Obuchowski, N. 1998. "Sample Size Calculations in Studies of Test Accuracy." *Statistical Methods in Medical Research*, 7, pages 371-392.

Obuchowski, N. and McClish, D. 1997. "Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices." *Statistics in Medicine*, 16, pages 1529-1542.

Odeh, R.E. and Fox, M. 1991. *Sample Size Choice*. Marcel Dekker, Inc. New York, NY.

O'Neill and Wetherill. 1971 "The Present State of Multiple Comparison Methods," *The Journal of the Royal Statistical Society*, Series B, vol.33, 218-250).

Ostle, B. and Malone, L. C. 1988. *Statistics in Research. Fourth Edition*. Iowa State Press. Ames, Iowa. A comprehension book on statistical methods.

Owen, Donald B. 1956. "Tables for Computing Bivariate Normal Probabilities," *Annals of Mathematical Statistics*, Volume 27, pages 1075-1090.

Owen, Donald B. 1965. "A Special Case of a Bivariate Non-Central t-Distribution," *Biometrika*, Volume 52, pages 437-446.

P

Parmar, M.K.B. and Machin, D. 1995. *Survival Analysis*. John Wiley and Sons. New York.

Pan, Z. and Kupper, L. 1999. "Sample Size Determination for Multiple Comparison Studies Treating Confidence Interval Width as Random." *Statistics in Medicine* 18, 1475-1488.

Pearson, E.S. and Hartley, H.O. 1976. *Biometrika Tables For Statistics, Volume 1*. Biometrika Trust. London.

Phillips, Kem F. 1990. "Power of the Two One-Sided Tests Procedure in Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 18, No. 2, pages 137-144.

Piantadosi, S. 2005. *Clinical Trials – A Methodological Perspective*. John Wiley & Sons. New Jersey.

Pocock, S.J. 1977. "Group sequential methods in the design and analysis of clinical trials." *Biometrika*, 64, pages 191-199.

Pocock, S.J. 1983. *Clinical Trials – A Practical Approach*. John Wiley & Sons. New York.

Price, K., Storn R., and Lampinen, J. 2005. *Differential Evolution – A Practical Approach to Global Optimization.* Springer. Berlin, Germany.

Prihoda, Tom. 1983. "Convenient Power Analysis For Complex Analysis of Variance Models." *Poster Session of the American Statistical Association Joint Statistical Meetings*, August 15-18, 1983, Toronto, Canada. Tom is currently at the University of Texas Health Science Center. This article includes FORTRAN code for performing power analysis.

R

Ramsey, Philip H. 1978 “Power Differences Between Pairwise Multiple Comparisons,*” JASA*, vol. 73, no. 363, pages 479-485.

Rao, C.R. , Mitra, S.K., & Matthai, A. 1966. *Formulae and Tables for Statistical Work*. Statistical Publishing Society, Indian Statistical Institute, Calcutta, India.

Reboussin, D.M., DeMets, D.L., Kim, K, and Lan, K.K.G. 1992. “Programs for computing group sequential boundaries using the Lan-DeMets Method.” Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. John Wiley. New York, New York. This book provides a comprehensive mixture of theoretical and applied results in multivariate analysis. My evaluation may be biased since Al Rencher took me fishing when I was his student.

Robins, Greenland, and Breslow. 1986. “A General Estimator for the Variance of the Mantel-Haenszel Odds Ratio,*” American Journal of Epidemiology*, vol.42, pages 719-723.

Robins, Breslow, and Greenland. 1986. “Estimators of the Mantel-Haenszel variance consisten in both sparse data and large-strata limiting models,*” Biometrics*, vol. 42, pages 311-323.

Rosenberger, W.F., and Lachin, J.M. 2002. *Randomization in Clinical Trials – Theory and Practice*. John Wiley & Sons. New York.

S

Sachs, Lothar. 1984. *Applied Statistics: A Handbook of Techniques*. Springer-Verlag. New York, New York.

Sahai, Hardeo & Khurshid, Anwer. 1995. *Statistics in Epidemiology*. CRC Press. Boca Raton, Florida.

Schilling, Edward. 1982. *Acceptance Sampling in Quality Control*. Marcel-Dekker. New York.

Schlesselman, Jim. 1981. *Case-Control Studies*. Oxford University Press. New York. This presents a complete overview of case-control studies. It was our primary source for the Mantel-Haenszel test.

Schoenfeld, David A. 1983. “Sample-Size Formula for the Proportional-Hazards Regression Model” *Biometrics*, Volume 39, pages 499-503.

Schoenfeld, David A. and **Richter, Jane R.** 1982**.** “Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint,” *Biometrics*, March 1982, Volume 38, pages 163-170.

Schork, M. and Williams, G. 1980. "Number of Observations Required for the Comparison of Two Correlated Proportions." *Communications in Statistics-Simula. Computa.*, B9(4), 349-357.

Schuirmann, Donald. 1981**.** “On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval,” *Biometrics*, Volume 37, pages 617.

Schuirmann, Donald. 1987**.** “A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability,” *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 15, Number 6, pages 657-680.

Senn, Stephen. 1993. *Cross-over Trials in Clinical Research*. John Wiley & Sons. New York.

Senn, Stephen. 2002. *Cross-over Trials in Clinical Research*. Second Edition. John Wiley & Sons. New York.

Shuster, Jonathan J. 1990. *CRC Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton, Florida. This is an expensive book ($300) of tables for running log-rank tests. It is well documented, but at this price it better be.

Signorini, David. 1991. “Sample size for Poisson regression,” *Biometrika,* Volume 78, 2, pages 446-450.

Simon, Richard. “Optimal Two-Stage Designs for Phase II Clinical Trials,” *Controlled Clinical Trials,* 1989, Volume 10, pages 1-10.

Smith, R.L. 1984. “Sequential Treatment Allocation using Biased Coin Designs.” *Journal of the Royal Statistical Society B*. Volume 46, pages 519-543.

Stekel, D. 2003. *Microarray Bioinformatics.* Cambridge University Press. Cambridge, United Kingdom.

Statxact 5. 2001. *Statistical Software for exact nonparametric inference, user manual.* Cytel Software Corporation. Cambridge, Massachusetts.

Swets, John A. 1996. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics - Collected Papers.* Lawrence Erlbaum Associates. Mahway, New Jersey.

T

Tango, Toshiro. 1998. "Equivalence Test and Confidence Interval for the Difference in Proportions for the Paired-Sample Design." *Statistics in Medicine*, Volume 17, 891-908.

Therneau, T.M. and Grambsch, P.M. 2000. *Modeling Survival Data*. Springer: New York, New York. A the time of the writing of the Cox regression procedure, this book provides a thorough, up-to-date discussion of this procedure as well as many extensions to it. Recommended, especially to those with at least a masters in statistics.

Thode, Henry C. 2002. *Testing for Normality*. Marcel Dekker, Inc. New York.

Thompson, Simon G. 1998. *Encyclopedia of Biostatistics, Volume 4*. John Wiley & Sons. New York. Article on Meta-Analysis on pages 2570-2579.

Tubert-Bitter, P., Manfredi,R., Lellouch, J., Begaud, B. 2000. “Sample size calculations for risk equivalence testing in pharmacoepidemiology.” *Journal of Clinical Epidemiology* 53, 1268-1274.

Tukey, J.W. and McLaughlin, D.H. 1963. “Less Vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization.” *Sankhya, Series A* 25, 331-352.

U

Upton, G.J.G. 1982."A Comparison of Alternative Tests for the 2 x 2 Comparative Trial.", *Journal of the Royal Statistical Society,* Series A,, Volume 145, pages 86-105.

W

Walter, S.D., Eliasziw, M., and Donner, A. 1998. "Sample Size and Optimal Designs For Reliability Studies.” *Statistics in Medicine*, 17, 101-110.

Wei, L.J., and Lachin, J.M. 1988. “Properties of the Urn Randomization in Clinical Trials.” *Controlled Clinical Trials*. Volume 9, pages 345-364.

Welch, B.L. 1938. "The significance of the difference between two means when the population variances are unequal." *Biometrika*, 29, 350-362.

Westlake, W.J. 1981. “Bioequivalence testing—a need to rethink,” *Biometrics*, Volume 37, pages 591-593.

Whittemore, Alice. 1981. “Sample Size for Logistic Regression with Small Response Probability,” *Journal of the American Statistical Association*, Volume 76, pages 27-32.

Wilson, E.B.. 1927. “Probable Inference, the Law of Succession, and Statistical Inference,” *Journal of the American Statistical Association*, Volume 22, pages 209-212. This article discusses the ‘score’ method that has become popular when dealing with proportions.

Winer, B.J. 1991**.** *Statistical Principles in Experimental Design (Third Edition)*. McGraw-Hill. New York, NY. A very complete analysis of variance book.

Wolfinger, R., Tobias, R. and Sall, J. 1994. "Computing Gaussian likelihoods and their derivatives for general linear mixed models,” *SIAM Journal of Scientific Computing*, 15, no.6, pages 1294-1310.

Woolson, R.F., Bean, J.A., and Rojas, P.B. 1986. “Sample Size for Case-Control Studies Using Cochran’s Statistic,” *Biometrics*, Volume 42, pages 927-932.

Y

Yateman, Nigel A. and Skene, Allan M. 1992. "Sample Sizes for Proportional Hazards Survival Studies with Arbitrary Patient Entry and Loss to Follow-Up Distributions." *Statistics in Medicine*, 11:1103-1113.

Yuen, K.K. and Dixon, W. J. 1973. “The approximate behavior and performance of the two-sample trimmed t,” *Biometrika*, Volume 60, pages 369-374.

Yuen, K.K. 1974. “The two-sample trimmed t for unequal population variances,” *Biometrika*, Volume 61, pages 165-170.

Z

Zar, Jerrold H. 1984**.** *Biostatistical Analysis (Second Edition)*. Prentice-Hall. Englewood Cliffs, New Jersey. This introductory book presents a nice blend of theory, methods, and examples for a long list of topics of interest in biostatistical work.

Zhou, X., Obuchowski, N., McClish, D. 2002**.** *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc. New York, New York. This is a great book on the designing and analyzing diagnostic tests. It is especially useful for its presentation of ROC curves.

Index

Index entries are of the form “chapter-page”. A list of chapters is given in the Table of Contents.