User's Guide - II

One Mean, Two Means, and Cross-Over Designs

PASS
Power Analysis and Sample Size System

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

PASS User's Guide - II

Dr. Jerry L. Hintze
Kaysville, Utah 84037

Direct inquiries to:

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

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

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. Purchasing the 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 or 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 PASS 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 fifteen 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 (98, 2000, ME, NT, XP, 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.

Table of Contents

Chapter 400

One Mean

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 $\Pr(t > T_a) = \alpha$. 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*.

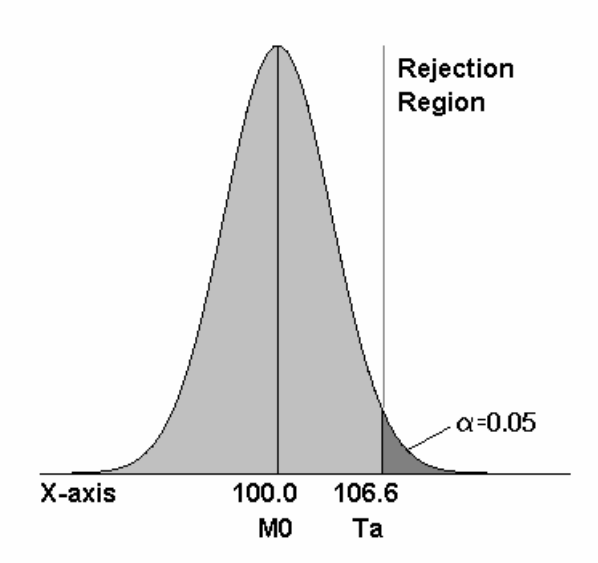
Following is a specific example. Suppose we want to test the hypothesis that a variable, *X*, has a mean of 100 versus the alternative hypothesis that the mean is greater than 100. Suppose that previous studies have shown that the standard deviation, σ, is 40. A random sample of 100 individuals is used.

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.

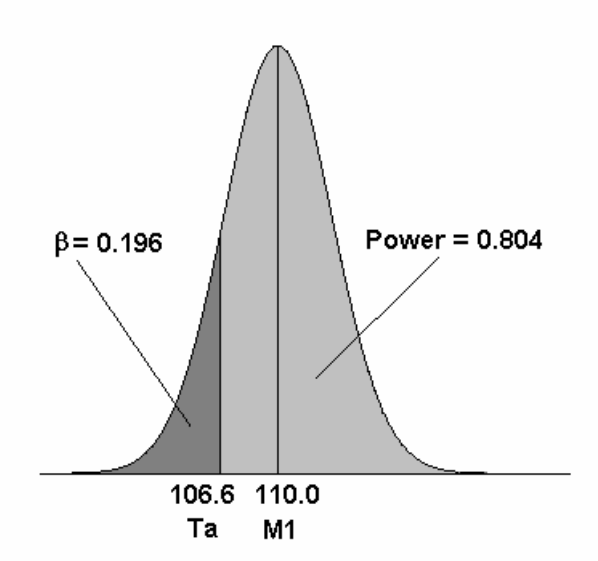
The *power* is the probability of rejecting H_0 when the true mean is 110. Since we reject H_0 when the calculated mean is greater than 106.6, the probability of a Type-II error (called β) is given by the dark, shaded area of the second graph. This value is 0.196. The power is equal to 1 - β or 0.804.

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 $1-\Phi(z_\alpha)=\alpha$, where $\Phi(x)$ is the area under the standardized normal curve to the left of *x*.
2. Calculate: $X_a = M0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ 1
3. Calculate: $z_a = \frac{X_a - M1}{\frac{\sigma}{\sqrt{n}}}$ 2
4. Power = $1-\Phi(z_a)$.

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 $1-T_{df}(t_\alpha)=\alpha$, where $T_{df}(t_\alpha)$ is the area under a central-*t* curve to the left of *x* and *df = n - 1*.
2. Calculate: $x_a = M0 + t_\alpha \frac{\sigma}{\sqrt{n}}$. 3
3. Calculate the noncentrality parameter: $\lambda = \frac{M1 - M0}{\frac{\sigma}{\sqrt{n}}}$. 4
4. Calculate: $t_a = \frac{x_a - M1}{\frac{\sigma}{\sqrt{n}}} + \lambda$ 5
5. Calculate: Power = $1-T'_{df,\lambda}(t_a)$, 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*.

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 that you will be concerned with.

Find

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

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

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

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.

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*.

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.

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*.

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). 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 was used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. Note that you should pick a value for alpha that represents the risk of a type-I error you are willing to take.

Experiment with different values between 0.01 and 0.10 to understand the relationship between alpha, beta, and sample size.

Beta (1 - Power)

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 equal means when in fact the means are different. You cannot make both a type-I and a type-II error in a single hypothesis test.

Values must be between zero and one. Historically, the value of 0.20 was used for beta. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as one minus beta. 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.

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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean0	**100**
Mean1	**110**
N	**20 to 120 by 20**
Standard Deviation	**40**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Alpha	**0.01 0.05 0.10**
Beta	*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 One-Sample T Test
Null Hypothesis: Mean0=Mean1 Alternative Hypothesis: Mean0<>Mean1
Unknown standard deviation.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect 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

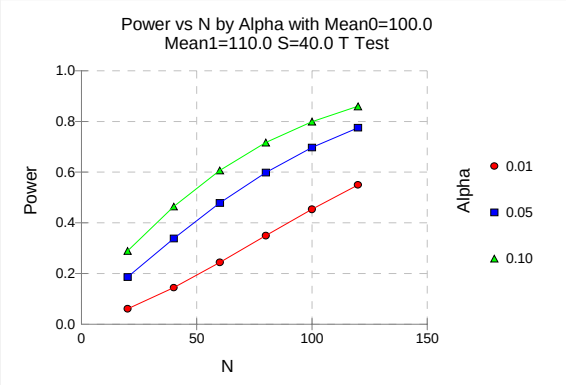
Report Definitions
Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population. To conserve resources, it should be small.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Mean0	**3300**
Mean1	**2475 2970 3135**
N	*Ignored since this is the Find setting*
Standard Deviation	**663**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	**0.10 0.20**

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
The standard deviation was assumed to be unknown.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Mean1 (Search<Mean0)**
Mean0	**3300**
Mean1	*Ignored since this is the Find setting*
N	**50**
Standard Deviation	**663**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	**0.20**

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
The standard deviation was assumed to be unknown.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Mean0	**0**
Mean1	**-5**
N	*Ignored since this is the Find setting.*
Standard Deviation	**10 12.5 15**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Alpha	**0.01 0.05**
Beta	**0.20**

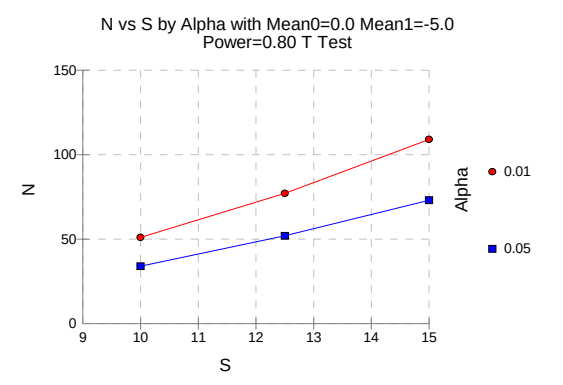
Annotated Output

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
The standard deviation was assumed to be unknown.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect 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. 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	**N**
Mean0	**0**
Mean1	**-5**
N	*Ignored since this is the Find setting.*
Standard Deviation	**10 12.5 15**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Double Exponential**
Alpha	**0.01 0.05**
Beta	**0.20**

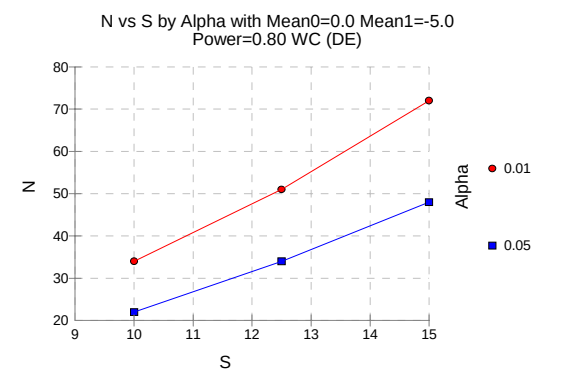
Annotated Output

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
The standard deviation was assumed to be unknown.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect 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. You can make these changes directly on your screen or you can load the template entitled Example6 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean0	**0**
Mean1	**1**
N	**12**
Standard Deviation	**1.25**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	*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 One-Sample T Test
Null Hypothesis: Mean0=Mean1 Alternative Hypothesis: Mean0<>Mean1
The standard deviation was assumed to be unknown.

Power	N	Alpha	Beta	Mean0	Mean1	S	Effect 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. You can make these changes directly on your screen or you can load the template entitled Example7 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Mean0	**0**
Mean1	**0.2**
N	*Ignored since this is the Find setting*
Standard Deviation	**1.0**
Known Standard Deviation	***Unchecked***
Population Size	**Infinite**
Alternative Hypothesis	**Ha: Mean0 <> Mean1**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	**0.20**

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
The standard deviation was assumed to be unknown.

Power	N	Alpha	Beta	Mean0	Mean1	Effect 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

Exponential Mean Test

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

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

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

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

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

$$E(t_0) = \begin{cases} \theta\sum_{i=1}^{r_0}\frac{1}{n-i+1} & \text{without replacement} \\ \frac{\theta r_0}{n} & \text{with replacement} \end{cases}$$

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

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

and

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

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

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

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

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

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

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

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 $r > 0$.

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

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

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

and

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

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 $p > 0.5$

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

where

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

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

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.

Data Tab

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

Find

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 *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 *Beta* when you want to calculate the power of an experiment or test.

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

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

so that

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

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

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

so that

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

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.

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.

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.

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.

Termination Criterion

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

1. **Fixed failures (*r*)**: terminate after *r* failures occur. This is also called *failure terminated* or *Type-II Censoring*.
2. **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).

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.

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.

Beta (Consumer's Risk)

This option specifies one or more values for the probability of a type-II error (beta), the consumer's risk. 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 0.20 was used for beta. Now, 0.10 is more popular. You should pick a value for beta that represents the risk of a type-II error you are willing to take.

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.

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.

Example1 - 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Theta0	**2**
Theta1	**1**
N	**5 to 50 by 5**
t0	**1**
Alternative Hypothesis	**Ha: Theta0 > Theta1**
Termination Criterion	**Fixed Time using Theta-hat**
Replacement Method	**Without Replacement**
Alpha	**0.01 0.05**
Beta	*Ignored since this is the Find setting*

Annotated Output

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

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.

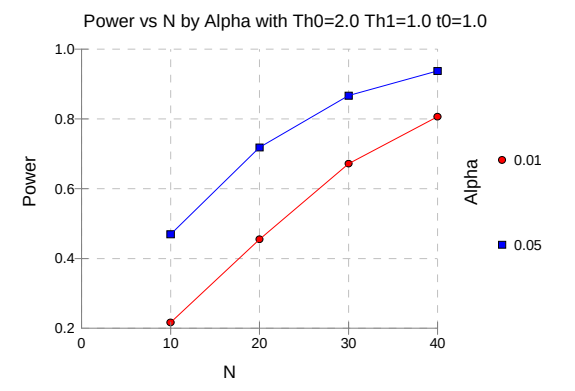
Power	N	Time t0	Theta0	Theta1	Target Alpha	Actual Alpha	Target Beta	Actual Beta	Theta C
0.21695	10	1.000	2.0	1.0	0.01000	0.01000		0.78305	0.7
0.45485	20	1.000	2.0	1.0	0.01000	0.01000		0.54515	1.0
0.67159	30	1.000	2.0	1.0	0.01000	0.01000		0.32841	1.1
0.80628	40	1.000	2.0	1.0	0.01000	0.01000		0.19372	1.2
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.

Plot Section



Example2 - 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**r**
Theta0	**5**
Theta1	**2**
N	**20 (this value is ignored)**
t0	**1**
Alternative Hypothesis	**Ha: Theta0 > Theta1**
Termination Criterion	**Fixed Failures, Fixed E(t0)**
Replacement Method	**Without Replacement**
Alpha	**0.01 0.05 0.10**
Beta	**0.05**

Annotated Output

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

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.

Power	r0 / N	Time E(t0)	Theta0	Theta1	Target Alpha	Actual Alpha	Target Beta	Actual 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

Tests of One Mean using Simulation

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 another ***PASS*** chapter 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

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, $M0$, 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 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 Procedure Templates.

Data Tab

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

Find

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.

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.

H1 (Alternative)

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.

Test Statistic

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). 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. Commonly, the value of 0.05 is used for two-tailed tests and 0.025 is used for one-tailed tests.

Beta (1 - Power)

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 equal means when in fact the means are different. One cannot make both a type-I and a type-II error in a single hypothesis test.

Values must be between zero and one. Historically, the value of 0.20 was used for beta. Now, 0.10 is more common. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as 1-beta. Hence, specifying beta also specifies the power. 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.

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.

M0 (Mean under 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 under 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.

Options Tab

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

Maximum Iterations

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

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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**1000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**T-Test**
N	**5 to 40 by 5**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Distribution Assuming H0	**N(M0 S)**
Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**1**
S	**2**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**
Show Definitions	**Checked**
Show Plots	**Checked**
Summary Statement Rows	**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

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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]		

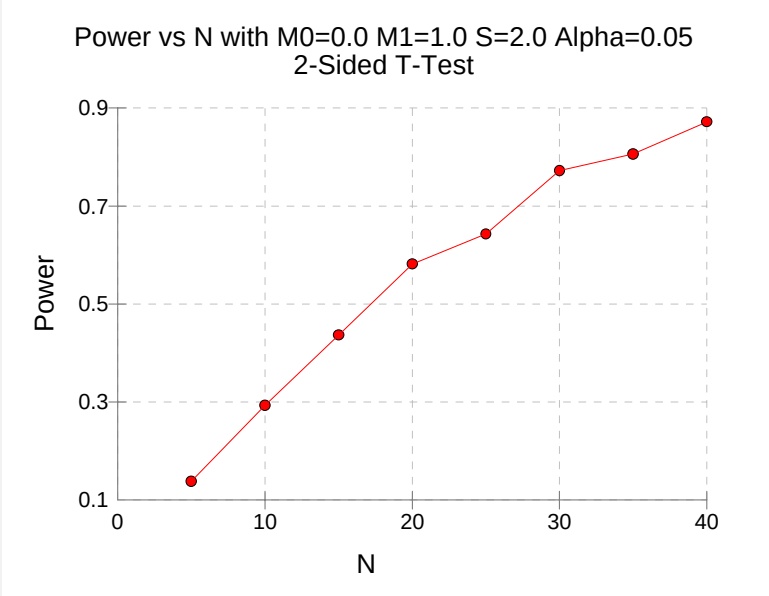
Notes:
Second Row: (Power Inc.) [95% LCL and UCL Power] (Alpha Inc.) [95% LCL and UCL Alpha]
Number of Monte Carlo Samples: 1000. Simulation Run Time: 17.81 seconds.

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



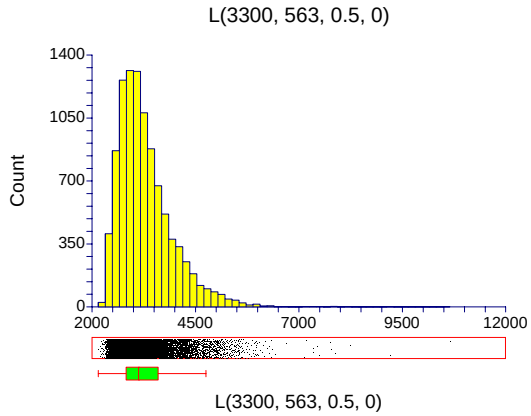
This plot shows the relationship between sample size and power.

Example2 - 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Simulations	**1000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**T-Test**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.20**
Distribution Assuming H0	**L(M0 S G 0)**
Distribution Assuming H1	**L(M1 S G 0)**
M0 (Mean under H0)	**2475**
M1 (Mean under 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**

Annotated Output

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

Numeric Results of Search for N

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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!

Example3 – 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**1000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**T-Test**
N	**6**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Distribution Assuming H0	**L(M0 S G 0)**
Distribution Assuming H1	**L(M1 S G 0)**
M0 (Mean under H0)	**2475**
M1 (Mean under 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**

Annotated Output

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

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)

N	H0 Mean (Mean0)	H1 Mean (Mean1)	Target Alpha	T-Test Power	Wilcoxon Power	Sign 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

N	H0 Mean (Mean0)	H1 Mean (Mean1)	Target Alpha	T-Test Alpha	Wilcoxon Alpha	Sign 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.

Example4 - 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**10000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**T-Test**
N	**12**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Distribution Assuming H0	**N(M0 S)**
Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under 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**

Annotated Output

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

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.

Example5 - 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. 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	**N**
Simulations	**200**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**T-Test**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.20**
Distribution Assuming H0	**N(M0 S)**
Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**0.20**
S	**1**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**

Annotated Output

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

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

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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.

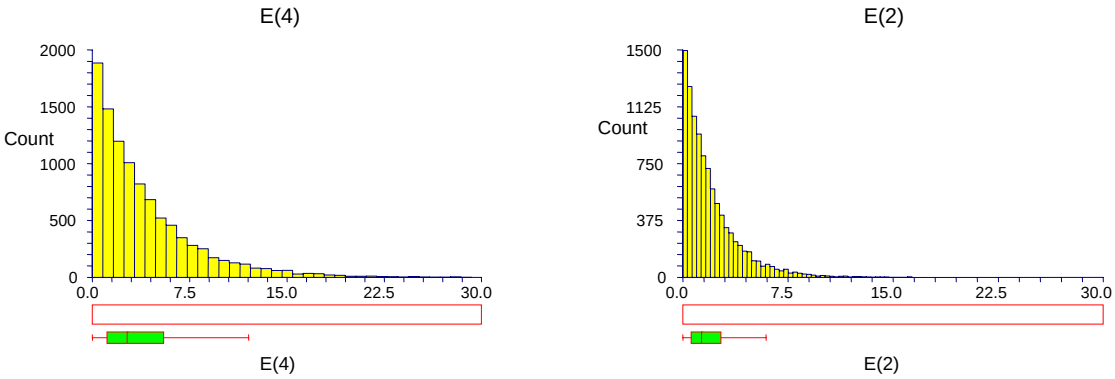
Note that using a simulation size of only 200, the estimated sample size of 211 is still close to the exact value of 199. We ran this simulation several times and obtained sample sizes between 187 and 211.

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

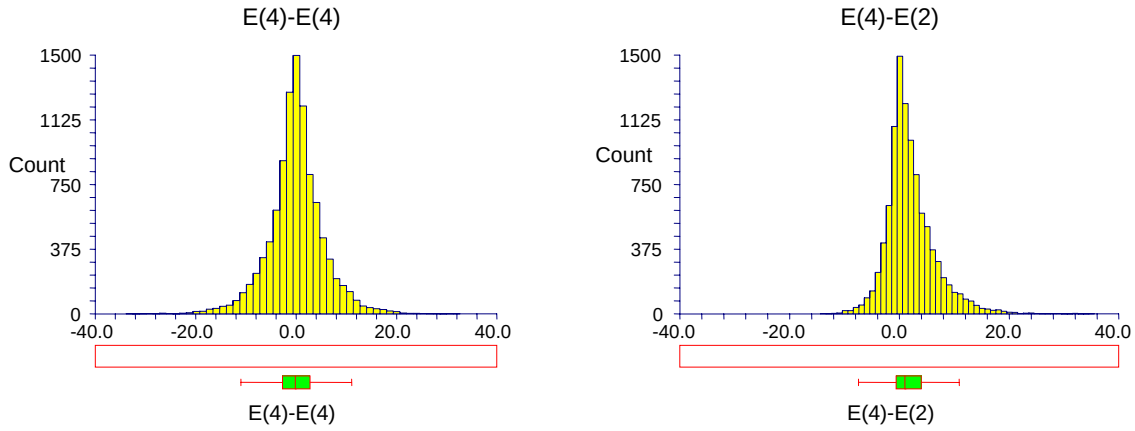
Example6 – Power of the Wilcoxon Test

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. You can make these changes directly on your screen or you can load the template entitled Example6 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**5000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**Wilcoxon**
N	**10 20 30 40 50**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Distribution Assuming H0	**E(M0)-E(M0)**
Distribution Assuming H1	**E(M0)-E(M1)**
M0 (Mean under H0)	**4**
M1 (Mean under H1)	**2**
Reports Tab	
Show Numeric Report	**Checked**
Show Inc's & 95% C.I.'s	**Checked**

Annotated Output

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

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

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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.

Example7 – 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. You can make these changes directly on your screen or you can load the template entitled Example7 by clicking the Template tab and loading this template.

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

Annotated Output

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

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

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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.

Example8 – 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. You can make these changes directly on your screen or you can load the template entitled Example8 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**1000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**Bootstrap**
N	**20**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Distribution Assuming H0	**N(M0 S)**
Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**155**
M1 (Mean under H1)	**160 to 190 by 10**
S	**33**
Options Tab	
Bootstrap Iterations	**100**

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

Annotated Output

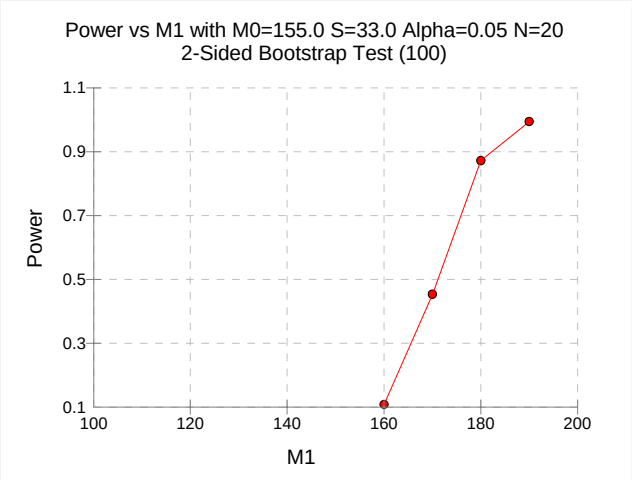
Click the Run button to perform the calculations and generate the following output.

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)

Power	N	H0 Mean0	H1 Mean1	Target Alpha	Actual 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)

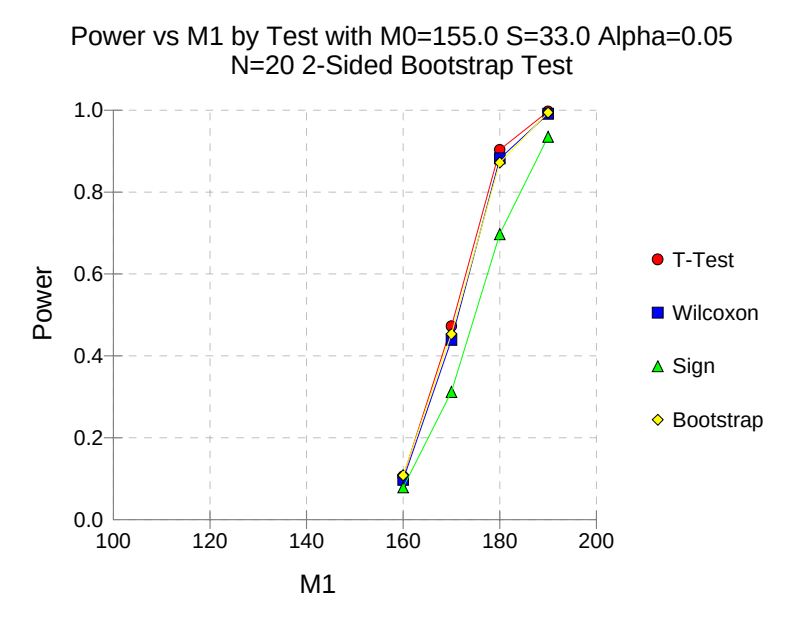
N	H0 Mean (Mean0)	H1 Mean (Mean1)	Target Alpha	T-Test Power	Wilcoxon Power	Sign Power	Bootstrap Power
20	155.0	160.0	0.050	0.105	0.097	0.078	0.108
20	155.0	170.0	0.050	0.472	0.439	0.312	0.453
20	155.0	180.0	0.050	0.903	0.882	0.697	0.872
20	155.0	190.0	0.050	0.997	0.991	0.935	0.994

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)

N	H0 Mean (Mean0)	H1 Mean (Mean1)	Target Alpha	T-Test Alpha	Wilcoxon Alpha	Sign Alpha	Bootstrap 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.

Example9 – 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	**Power**
Simulations	**1000**
H1 (Alternative)	**Mean<>M0**
Test Statistic	**T-Test**
N	**20 to 60 by 20**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Distribution Assuming H0	**E(M0)**
Distribution Assuming H1	**E(M1)**
M0 (Mean under H0)	**10**
M1 (Mean under H1)	**20**
Options Tab	
Bootstrap Iterations	**100**
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**

Annotated Output

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

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

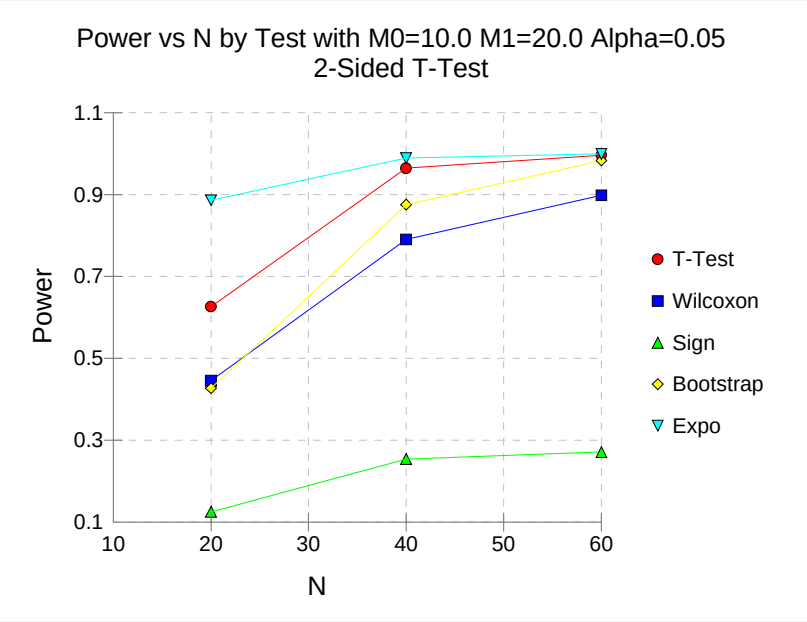
N	H0 Mean (Mean0)	H1 Mean (Mean1)	Target Alpha	T-Test Power	Wilcoxon Power	Sign Power	Bootstrap Power	Expo Power
20	10.0	20.0	0.050	0.626	0.445	0.125	0.427	0.886
40	10.0	20.0	0.050	0.964	0.790	0.254	0.875	0.989
60	10.0	20.0	0.050	0.996	0.898	0.271	0.983	0.999

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)

N	H0 Mean (Mean0)	H1 Mean (Mean1)	Target Alpha	T-Test Alpha	Wilcoxon Alpha	Sign Alpha	Bootstrap Alpha	Expo 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 Tests of 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 One-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 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

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

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.*

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

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

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

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

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

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 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

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

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

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

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 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

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

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

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

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 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

$$H_0: \mu_T \geq \mu_R - |\varepsilon| \quad \text{versus} \quad H_1: \mu_T < \mu_R - |\varepsilon|$$

$$H_0: \mu_T - \mu_R \geq -|\varepsilon| \quad \text{versus} \quad H_1: \mu_T - \mu_R < -|\varepsilon|$$

$$H_0: \delta \geq -|\varepsilon| \quad \text{versus} \quad H_1: \delta < -|\varepsilon|$$

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 $D0$ 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, $D0$, 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

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

   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 $1 - T_{n-1}(t_\alpha) = \alpha$, where $T_{n-1}(t_\alpha)$ 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 $t_a = \frac{x_a - D1}{\frac{\sigma}{\sqrt{n}}} + \lambda$

5. Calculate the power = $1 - T'_{n-1,\lambda}(t_a)$, where $T'_{n-1,\lambda}(x)$ 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 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.

Data Tab

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

Find

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

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

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

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.

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.

|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.

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*.

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject a true H0. Since this is a one-sided test, the value of 0.025 is commonly used for alpha.

Beta (1 - Power)

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 you should. Values must be between zero and one. The value of 0.10 is recommended for beta.

Power is defined as one minus beta. 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.

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*.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

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

Annotated Output

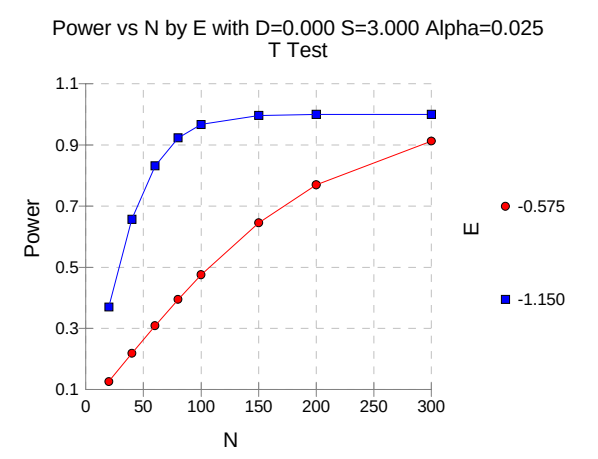
Click the Run button to perform the calculations and generate the following output.

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

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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.

Example2 - 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
N	*Ignored since this is the Find setting*
S (Std Deviation)	**3**
Alpha	**0.025**
Beta	**0.10**
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 Non-Inferiority Test (H0: D <= -\|E\|; H1: D > -\|E\|)
Test Statistic: T-Test

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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|.

Example3 - 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.5**
D (True Difference)	**0.5**
N	*Ignored since this is the Find setting*
S (Std Deviation)	**1**
Alpha	**0.05**
Beta	**0.20**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

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

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

PASS has also obtained a sample size of 8.

Example4 - 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**10**
D (True Difference)	**0 2**
N	*Ignored since this is the Find setting*
S (Std Deviation)	**28.284271**
Alpha	**0.025**
Beta	**0.10**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

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

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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.

Example5 - 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	**N**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.2**
D (True Difference)	**-.1**
N	*Ignored since this is the Find setting*
S (Std Deviation)	**.2**
Alpha	**0.05**
Beta	**0.20**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

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

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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 Interval for the Mean

Introduction

This routine calculates the sample size necessary to achieve a required precision at a stated confidence coefficient for a confidence interval about the mean when the underlying data distribution is normal.

Technical Details

For a single mean from a normal distribution, a two-sided, $100(1-\alpha)\%$ confidence interval is calculated by

$$\bar{x} \pm \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}$$

when σ^2 (the variance of *X*) is known or by

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

when σ^2 (the variance of *X*) is unknown.

Notice that the confidence interval is calculated using an estimate of the mean plus or minus a quantity that represents the precision (or margin of error) of this estimate. That is, the width of the confidence interval is equal to twice the term on the right of the expression. We will label this half-width, *D*, and call it the precision of the confidence interval.

The basic equation for determining sample size 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. When the standard deviation is unknown, we must supply an estimate of that value.

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 Coefficient

The confidence coefficient, $1 - \alpha$, 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 $1 - \alpha$.

Notice that is a long term statement about many, many samples.

Power

Notice that these formulas do not contain a statement about the power. In fact, since we are calculating confidence intervals and not conducting hypothesis tests, we cannot commit the errors that are possible with those tests. A natural question is, if we obtained a sample size based on the confidence interval formulas and then conducted a hypothesis test, what would be the power of the test? The answer is about 0.50.

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.

Data Tab

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

Find

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

Precision

This is half the width of the confidence interval. That is, the confidence interval is formed by taking the sample mean plus and minus this amount. The smaller this amount, the more narrow the interval.

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

Confidence Coefficient

The confidence coefficient, $1-\alpha$, 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 $1-\alpha$. In power analysis, we specify α. When dealing with confidence intervals, we specify $1-\alpha$.

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*.

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.

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.

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 must use the results of a pilot study, a previous study, or a ball park estimate based on the range (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 size is greater than fifty.

Options Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

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.

Example1 - Calculating Sample Size

Suppose a study is planned in which you want to construct a confidence interval for the mean that is no wider than 7 units. You set the confidence coefficient at 0.95. Previous studies have shown that the standard deviation is about 28. Instead of looking at a precision of just 7, you want to see the sample sizes for a range of values from 5 to 9.

Calculate the necessary sample size.

Setup

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Sample Size)**
Precision	**5 to 9 by 1**
Confidence Coefficient	**0.95, 0.99**
Population Size	**Infinite**
N (Sample Size)	**Ignored**
S (Standard Deviation)	**28**

Annotated Output

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

Numeric Results

Numeric Results

Precision	C.C. Confidence Coefficient	N Sample Size	S Standard Deviation
5.0	0.95000	123	28.0
5.0	0.99000	209	28.0
6.0	0.95000	87	28.0
6.0	0.99000	149	28.0
7.0	0.95000	64	28.0
7.0	0.99000	110	28.0
8.0	0.95000	50	28.0
8.0	0.99000	86	28.0
9.0	0.95000	40	28.0
8.9	0.99000	69	28.0

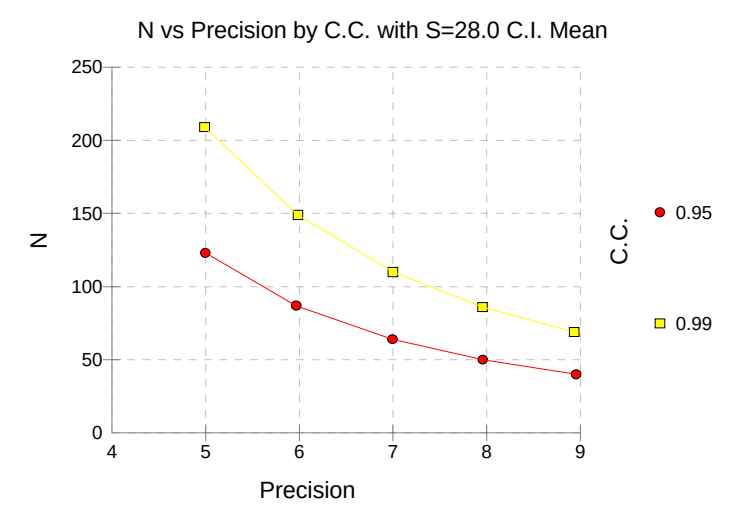
Unknown standard deviation.

Report Definitions
Precision is the plus and minus value used to create the confidence interval.
Confidence Coefficient is probability value associated with the confidence interval.
N is the size of the sample drawn from the population.
The standard deviation of the population measures the variability in the population.

Summary Statements
A sample size of 123 produces a 95% confidence interval equal to the sample mean plus or minus 5.0 when the estimated standard deviation is 28.0.

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

Plot Section



This plot shows the sample size versus the precision for the two confidence coefficients.

Example2 - 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Sample Size)**
Precision	**2**
Confidence Coefficient	**0.95**
Population Size	**Infinite**
N (Sample Size)	**Ignored**
S (Standard Deviation)	**3**
Known Standard Deviation	**Checked**

Annotated Output

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

Numeric Results

Numeric Results

Precision	C.C. Confidence Coefficient	N Sample Size	S Standard Deviation
2.0	0.95000	9	3.0

Known standard deviation.

PASS also calculated the necessary sample size to be 9.

Chapter 430

Two Means

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.

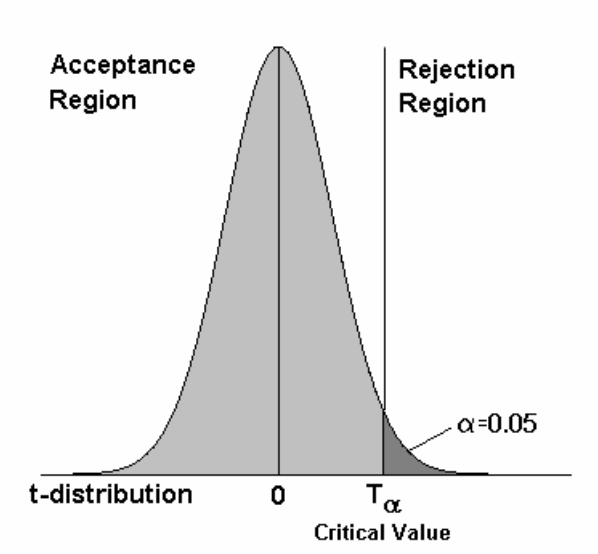
Test Procedure

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 $\Pr(T > T_\alpha) = \alpha$.

   Figure 1 - Find the Critical Value

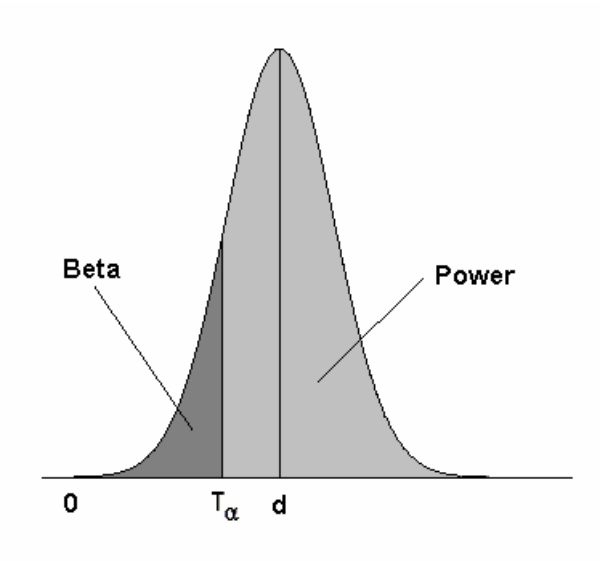
   Again, select elect T_α so that if the means of the two populations are equal, *t* statistics calculated from two samples drawn from those populations will only exceed T_α exactly $100\alpha\%$ of the time.

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.

Figure 2 - Computing the Power



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.

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: $z_p = \frac{z_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}}$
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: $t_p = \frac{t_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}} + \lambda$
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: $f = \frac{\sigma_{\bar{x}}^4}{\frac{\sigma_1^4}{N_1^2(N_1+1)} + \frac{\sigma_2^4}{N_2^2(N_2+1)}} - 2$

   which is the adjusted degrees of freedom. Often, this is rounded to the next highest integer.
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: $\lambda = \frac{d}{\sigma_{\bar{x}}}$,1 the noncentrality parameter.
5. Calculate: $t_p = \frac{t_\alpha \sigma_{\bar{x}} - d}{\sigma_{\bar{x}}} + \lambda$ 2
1. 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 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 that you will be concerned with.

Find

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*, *Beta*, *N1*, and *N2*. In most situations, you will select either *Beta* or *N1*.

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

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

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.

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*.

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

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). 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. The recommended range is 0.001 to 0.10. Historically, the value of 0.05 was 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.

Below are some values of alpha and corresponding odds of rejection.

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

Beta (1 - Power)

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 equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Lately, 0.10 is becoming popular. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

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.

Note that each value of the alternative hypothesis will have a different power.

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 Std 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.

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

Variable	Count	Mean	Standard Deviation	Standard Error	95% LCL of Mean	95% UCL 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 Hypothesis	*T* Value	Prob Level	Decision (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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta**
Mean1	**20.9**
Mean2	**17.8**
N1	**5 10 15 20 25 30 50**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Alpha	**0.01 0.05**
Beta	*Ignored since this is the Find setting*

S1.....**3.67**
S2.....**3.01**
Known Std Deviation.....***Not checked***
Axes 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.

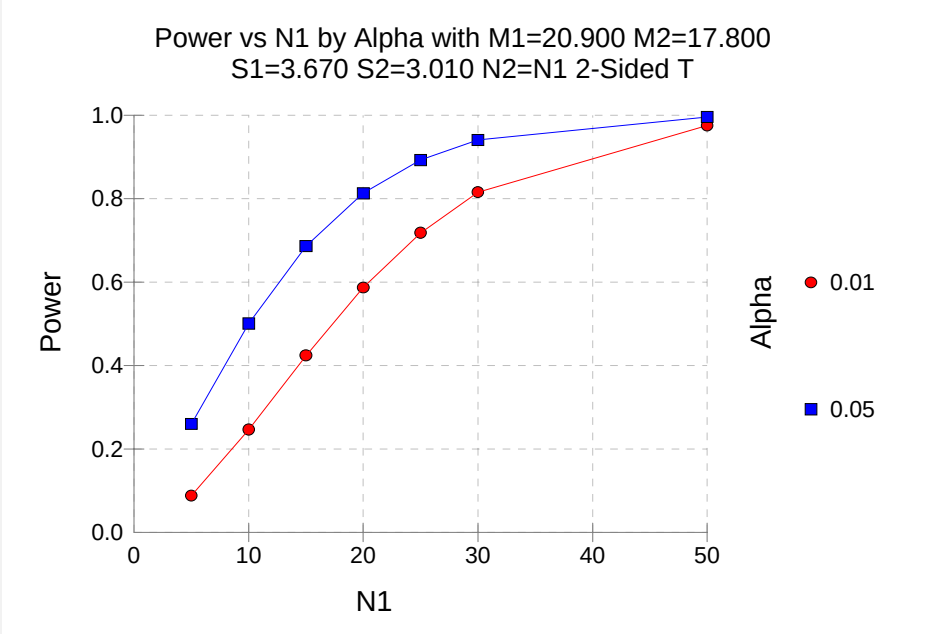
Power	N1	N2	Allocation Ratio	Alpha	Beta	Mean1	Mean2	S1	S2
0.08825	5	5	1.00	0.01000	0.91175	20.900	17.800	3.670	3.010
0.24642	10	10	1.00	0.01000	0.75358	20.900	17.800	3.670	3.010
0.42417	15	15	1.00	0.01000	0.57583	20.900	17.800	3.670	3.010
0.58661	20	20	1.00	0.01000	0.41339	20.900	17.800	3.670	3.010
0.71790	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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Mean1	**20.9**
Mean2	**17.8**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Alpha	**0.01 0.05**
Beta	**0.20**
S1	**3.67**
S2	**3.01**
Known Std Deviation	***Not checked***

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.

Power	N1	N2	Allocation 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Mean2 (Search>Mean1)**
Mean1	**0**
Mean2	*Ignored since this is the Find setting*
N1	**10**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	**0.20**
S1	**3.67**
S2	**3.01**
Known Std Deviation	***Not checked***

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.

Power	N1	N2	Allocation 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Mean1	**0**
Mean2	**15**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Alpha	**0.01 0.05**
Beta	**0.10**
S1	**10 12.5 15**
S2	**S1**
Known Std Deviation	***Not checked***

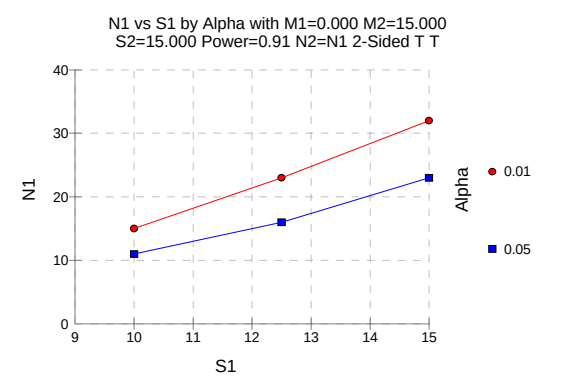
Annotated Output

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.

Power	N1	N2	Allocation 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. 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	**N1**
Mean1	**0**
Mean2	**15**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Double Exponential**
Alpha	**0.01 0.05**
Beta	**0.10**
S1	**10 12.5 15**
S2	**S1**
Known Std Deviation	***Not checked***

Annotated 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.

Power	N1	N2	Allocation 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. You can make these changes directly on your screen or you can load the template entitled Example6 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Mean1	**0**
Mean2	**5**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	**0.10**
S1	**10**
S2	**S1**
Known Std Deviation	***Not checked***

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.

Power	N1	N2	Allocation 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. You can make these changes directly on your screen or you can load the template entitled Example7 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean1	**0**
Mean2	**1**
N1	**15**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1 <> Mean2**
Nonparametric Adjustment	**Ignore**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
S1	**0.7206**
S2	**S1**
Known Std Deviation	***Not checked***

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.

Power	N1	N2	Allocation 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 435

Comparing Two Exponential Means

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

$$\hat{\theta}_i = \frac{\sum_{\text{over } j} t_{ij}}{r_i}, \quad i = 1,2$$

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 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.

Data Tab

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

Find

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

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

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

so that

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

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

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

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

so that

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

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

Note that only the ratio of theta1 and theta2 is used in the calculations.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha), also called the significance level. 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.

Beta (1 - Power)

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 equal thetas when in fact they are different.

Values must be between zero and one. Historically, the value of 0.20 was used for beta. Now, 0.10 is more popular. You should pick a value for beta that represents the risk of a type-II error that you are willing to take.

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.

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.

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*.

Example1 - 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Theta1	**1.3**
Theta2	**1.0**
Alternative Hypothesis	**Ha: Theta1 <> Theta2**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
N1	**5 20 50 100 200 300 400 500**
N2	**Use R**
R (Sample Allocation Ratio)	**1.0**

Annotated Output

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

Numeric Results

H0: Theta1 = Theta2. Ha: Theta1 <> Theta2.

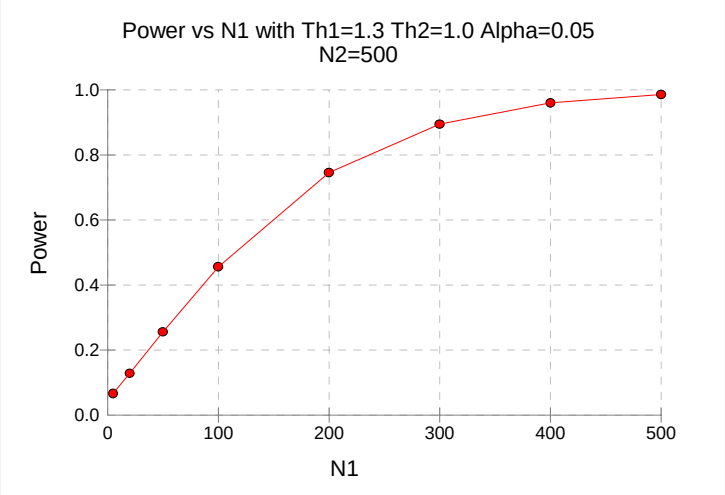
Power	N1	N2	Allocation Ratio	Alpha	Beta	Theta1	Theta2	Theta1/ Theta2
0.06652	5	5	1.00000	0.05000	0.93348	1.3	1.0	1.30000
0.12839	20	20	1.00000	0.05000	0.87161	1.3	1.0	1.30000
0.25602	50	50	1.00000	0.05000	0.74398	1.3	1.0	1.30000
0.45619	100	100	1.00000	0.05000	0.54381	1.3	1.0	1.30000
0.74551	200	200	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.

Plot Section



Example2 - 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Theta1	**1.3**
Theta2	**1.0**
Alternative Hypothesis	**Ha: Theta1 > Theta2**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
N1	**20**
N2	**Use R**
R (Sample Allocation Ratio)	**1.0**

Annotated Output

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

Numeric Results

H0: Theta1 = Theta2. Ha: Theta1 > Theta2.

Power	N1	N2	Allocation Ratio	Alpha	Beta	Theta1	Theta2	Theta1/ 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

Tests of Two Means Using Simulation

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). Tukey and Dixon (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

$$T_{tg} = \frac{\left(\overline{X}_{1tg} - \overline{X}_{2tg}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{SSD_{1wg} + SSD_{2wg}}{h_1 + h_2 - 2}\left(\frac{1}{h_1} + \frac{1}{h_2}\right)}}$$

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

$$\frac{1}{f} = \frac{c^2}{h_1 - 1} + \frac{1 - c^2}{h_2 - 1}$$

$$c = \frac{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)}}{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)} + \frac{SSD_{2wg}}{h_2\left(h_2 - 1\right)}}$$

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 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).

$$z = \frac{W_1 - \frac{N_1(N_1 + N_2 + 1)}{2} + C}{s_W}$$

where

$$W_1 = \sum_{k=1}^{N_1} Rank(X_{1k})$$

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 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.

Data Tab

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

Find

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.

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.

H1 (Alternative)

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**.

Test Statistic

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).

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*.

Group 1 (and 2) Dist'n | 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) Dist'n | 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.

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). 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. Commonly, the value of 0.05 is used for two-tailed tests and 0.025 is used for one-tailed (non-inferiority) tests.

Beta (1 - Power)

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 equal means when in fact the means are different. You cannot make both a type-I and a type-II error in a single hypothesis test.

Values must be between zero and one. Historically, the value of 0.20 was used for beta. Now, 0.10 is more common. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as one minus beta. Hence, specifying beta also specifies the power. 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.

Options Tab

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

Maximum Iterations

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 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.

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.

Reports Tab

The Reports tab contains settings about the format of the output.

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).

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.

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.

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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N1	**50 100 200**
N2	**Use R**
R	**1.0**
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 under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**
Alpha	**0.05**
Beta	*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 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

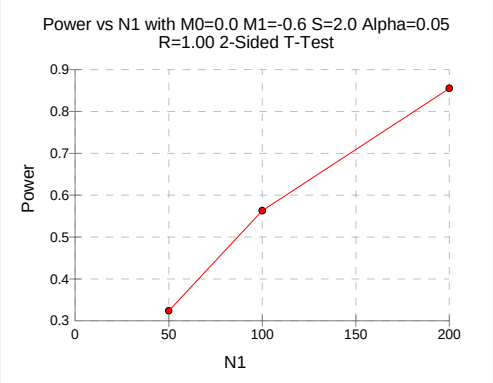
Power	N1/N2	H0 Diff0	H1 Diff1	Target Alpha	Actual Alpha	Beta	M0	M1	S
0.324	50/50	0.0	-0.6	0.050	0.056	0.676	0.0	0.6	2.0
(0.021)	[0.303	0.345]			(0.010)	[0.045	0.066]		
0.563	100/100	0.0	-0.6	0.050	0.047	0.437	0.0	0.6	2.0
(0.022)	[0.541	0.585]			(0.009)	[0.038	0.056]		
0.855	200/200	0.0	-0.6	0.050	0.045	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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
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 under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**
Alpha	**0.05**
Beta	**0.10**

Annotated Output

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

Numeric Results of Search for N

Power	N1/N2	H0 Diff0	H1 Diff1	Target Alpha	Actual 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N1	**50 100 200**
N2	**Use R**
R	**1.0**
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 under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

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

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)

N1/N2	H0 Diff (Diff0)	H1 Diff (Diff1)	Target Alpha	T-Test Power	Welch Power	Trim. T-Test Power	Trim. Welch Power	Mann Whit'y 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)

N1/N2	H0 Diff (Diff0)	H1 Diff (Diff1)	Target Alpha	T-Test Alpha	Welch Alpha	Trim. T-Test Alpha	Trim. Welch Alpha	Mann Whit'y 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

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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**10000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N1	**15**
N2	**Use R**
R	**1.0**
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 under H0)	**0**
M1 (Mean under H1)	**1**
S	**0.7206**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*

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

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

Power	N1/N2	H0 Diff0	H1 Diff1	Target Alpha	Actual 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. 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	**N1**
Simulations	**2000**
H1 (Alternative)	**Diff<Diff0**
Test Statistic	**T-Test**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
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 under H0)	**96**
M1 (Mean under H1)	**100**
S	**6**
Alpha	**0.025**
Beta	**0.10**

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

Power	N1/N2	H0 Diff0	H1 Diff1	Target Alpha	Actual 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. You can make these changes directly on your screen or you can load the template entitled Example6 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N1	**20**
N2	**Use R**
R	**1.0**
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 under H0)	**0**
M1 (Mean under H1)	**1**
S	**1**
A	**1 5 10**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

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

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]

N1/N2	H0 Diff (Diff0)	H1 Diff (Diff1)	Target Alpha	T-Test Power	Welch Power	Trim. T-Test Power	Trim. Welch Power	Mann Whit'y 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

N1/N2	H0 Diff (Diff0)	H1 Diff (Diff1)	Target Alpha	T-Test Alpha	Welch Alpha	Trim. T-Test Alpha	Trim. Welch Alpha	Mann Whit'y 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. You can make these changes directly on your screen or you can load the template entitled Example7 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N1	**20**
N2	**Use R**
R	**1.0**
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 under H0)	**0**
M1 (Mean under H1)	**1**
S	**1**
G	**0 0.5 0.9**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

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

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)

N1/N2	H0 Diff (Diff0)	H1 Diff (Diff1)	Target Alpha	T-Test Power	Welch Power	Trim. T-Test Power	Trim. Welch Power	Mann Whit'y 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

N1/N2	H0 Diff (Diff0)	H1 Diff (Diff1)	Target Alpha	T-Test Alpha	Welch Alpha	Trim. T-Test Alpha	Trim. Welch Alpha	Mann Whit'y 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

Test of the Ratio of Two Means

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

$$H_0: \phi = \phi_0$$

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 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

$$\begin{aligned}
\mu_Y &= \left(e^{\mu_X + \frac{\sigma_X^2}{2}}\right) \\
\sigma_Y^2 &= \mu_Y^2\left(e^{\sigma_X^2} - 1\right)
\end{aligned}$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Find

This option specifies the parameter to be solved for from the other parameters. In most situations, you will select either Beta for a power analysis or *N1* for sample size determination.

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*.

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.

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.

Alpha (Significance Level)

Specify one or more values of alpha, the probability of a type-I error which is rejecting the null hypothesis of equality when in fact the groups are different. Note that the valid range is 0 to 1, but typical values are between 0.01 and 0.20.

You can enter a range of values such as *0.05 0.10 0.15* or *0.5 to 0.15 by 0.05*.

Beta (1 - Power)

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 equal means when in fact the means are different.

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.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Alternative Hypothesis	**R1>R0 (One-Sided)**
R0	**1.0**
R1	**1.1 1.2**
COV	**1.2**
Alpha	**0.025**
Beta	*Ignored since this is the Find setting*
N1	**100 to 900 by 200**
N2	**Use R**
R	**1.0**

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.

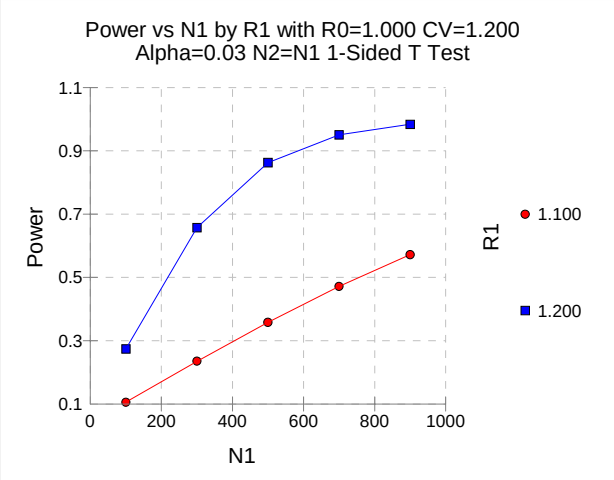
Power	Group Sample Sizes (N1/N2)	Mean Ratio Under H0 (R0)	Mean Ratio Under H1 (R1)	Effect Size (ES)	Coefficient of Variation (COV)	Significance Level (Alpha)	Beta
0.1057	100/100	1.000	1.100	0.1009	1.200	0.0250	0.8943
0.2351	300/300	1.000	1.100	0.1009	1.200	0.0250	0.7649
0.3581	500/500	1.000	1.100	0.1009	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.

Plot Section



This plot shows the power versus the sample size.

Example2 –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.

$$\begin{aligned} S1 &= \sqrt{\ln\left(COV^2 + 1\right)} \\ &= \sqrt{\ln\left(1.2^2 + 1\right)} \\ &= 0.944456 \end{aligned}$$

$$\begin{aligned} Mean2 &= \ln\left(R1\right) & Mean2 &= \ln\left(R1\right) \\ &= \ln\left(1.10\right) & &= \ln\left(1.20\right) \\ &= 0.095310 & &= 0.182322 \end{aligned}$$

Setup

Load the *PASS: Means: 2: Inequality: Differences* panel. You can enter the following parameter values or load Example1c.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean1	**0**
Mean2	**0.095310 0.182322**
N1	**100 to 900 by 200**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Ha: Mean1<Mean2**
Nonparametric Adjustment	**Ignore**
S1	**0.944456**
S2	**S1**
Alpha	**0.025**
Beta	*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 Two-Sample T-Test
Null Hypothesis: Mean1=Mean2. Alternative Hypothesis: Mean1<Mean2
The standard deviations were assumed to be unknown and equal.

Power	N1	N2	Allocation 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
0.2351	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
0.6571	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 Tests of the Difference in Two Means

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

$$\mathrm{H}_0: \mu_1 - \mu_2 \leq D \text{ versus } \mathrm{H}_1: \mu_1 - \mu_2 > D$$

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*.

$$\mathrm{H}_0: \mu_1 - \mu_2 \geq D \text{ versus } \mathrm{H}_1: \mu_1 - \mu_2 < D$$

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.

$H_0: \mu_1 \le \mu_2 - |\varepsilon|$ versus $H_1: \mu_1 > \mu_2 - |\varepsilon|$

$H_0: \mu_1 - \mu_2 \le -|\varepsilon|$ versus $H_1: \mu_1 - \mu_2 > -|\varepsilon|$

$H_0: \delta \le -|\varepsilon|$ versus $H_1: \delta > -|\varepsilon|$

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.

$H_0: \mu_1 \ge \mu_2 + |\varepsilon|$ versus $H_1: \mu_1 < \mu_2 + |\varepsilon|$

$H_0: \mu_1 - \mu_2 \ge |\varepsilon|$ versus $H_1: \mu_1 - \mu_2 < |\varepsilon|$

$H_0: \delta \ge |\varepsilon|$ versus $H_1: \delta < |\varepsilon|$

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.

$H_0: \mu_1 \le \mu_2 + |\varepsilon|$ versus $H_1: \mu_1 > \mu_2 + |\varepsilon|$

$H_0: \mu_1 - \mu_2 \le |\varepsilon|$ versus $H_1: \mu_1 - \mu_2 > |\varepsilon|$

$H_0: \delta \le |\varepsilon|$ versus $H_1: \delta > |\varepsilon|$

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.

$H_0: \mu_1 \geq \mu_2 - |\varepsilon|$ versus $H_1: \mu_1 < \mu_2 - |\varepsilon|$

$H_0: \mu_1 - \mu_2 \geq -|\varepsilon|$ versus $H_1: \mu_1 - \mu_2 < -|\varepsilon|$

$H_0: \delta \geq -|\varepsilon|$ versus $H_1: \delta < -|\varepsilon|$

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.

$$t_{df} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left|\varepsilon\right|}{s_{\overline{X}_1 - \overline{X}_2}}$$

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 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.

$$t_f^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left|\varepsilon\right|}{s^*_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s^*_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2}{N_1\left(N_1 - 1\right)}\right) + \left(\frac{\sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_2\left(N_2 - 1\right)}\right)}$$

$$f = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\left(N_1 - 1\right)} + \frac{s_2^4}{N_2^2\left(N_2 - 1\right)}}$$

$$s_1 = \sqrt{\left(\frac{\sum_{i=1}^{N_1} \left(X_{1i} - \overline{X}_1 \right)^2}{N_1 - 1} \right)}, s_2 = \sqrt{\left(\frac{\sum_{i=1}^{N_2} \left(X_{2i} - \overline{X}_2 \right)^2}{N_2 - 1} \right)}$$

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 $1 - T_{df}(t_\alpha) = \alpha$, 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: $\lambda = \frac{|\varepsilon| - \delta}{\sigma_{\bar{x}}}$

4. Calculate: Power = $1 - T'_{df,\lambda}(t_\alpha)$, 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 $1 - T_f(t_\alpha) = \alpha$, 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: $\lambda = \frac{|\varepsilon|}{\sigma_{\bar{x}}}$, the noncentrality parameter.

5. Calculate: Power = $1 - T'_{f,\lambda}(t_\alpha)$, 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 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*.

Data Tab

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

Find

This option specifies the parameter to be calculated from the values of the other parameters. Under most conditions, you would select either *Beta* or *N1*.

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

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

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.

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.

|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 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.

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject the null hypothesis of inferiority when in fact the mean is not non-inferior. Since this is a one-sided test, the value of 0.025 is commonly used for alpha.

Beta (1 - Power)

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 you should. Values must be between zero and one. The value of 0.10 is recommended for beta.

Power is defined as one minus beta. 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.

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.

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 function 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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta & Power**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
N1	**10 50 100 200 300 500 600 800**
S (Std Deviation)	**3**
Alpha	**0.025**
Beta	*Ignored since this is the Find setting*
Nonparametric Adjustment	**Ignore**
Reports Tab	
Mean Decimals	**3**

Annotated Output

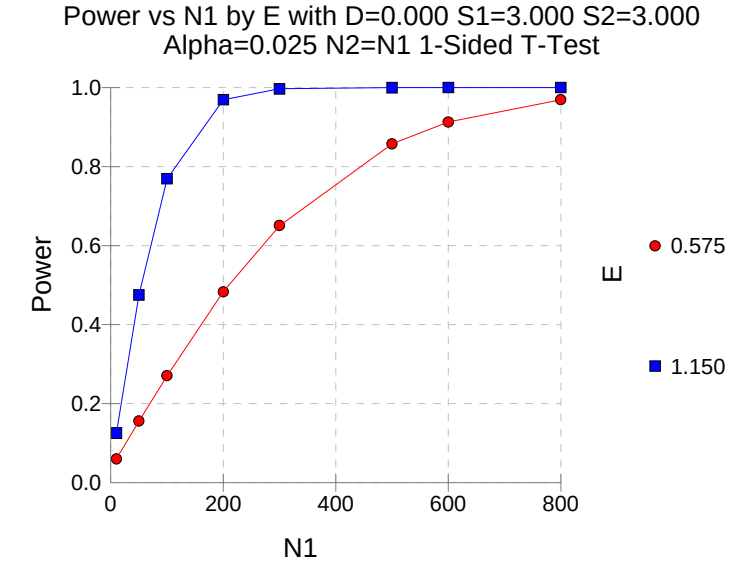
Click the Run button to perform the calculations and generate the following output.

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

Power	N1/N2	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation1 (SD1)	Standard Deviation2 (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.



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.

Example2 - Finding the Sample Size

Continuing with Example1, 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.575 1.15**
D (True Difference)	**0**
N1	*Ignored since this is the Find setting*
S (Std Deviation)	**3**
Alpha	**0.025**
Beta	**0.10**
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 for Non-Inferiority Test (H0: D <= -\|E\|; H1: D > -\|E\|)
Test Statistic: T-Test

Power	N1/N2	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation1 (SD1)	Standard Deviation2 (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|.

Example3 - 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**0.05**
D (True Difference)	**0**
N1	*Ignored since this is the Find setting*
S (Std Deviation)	**0.1**
Alpha	**0.05**
Beta	**0.20**
Nonparametric Adjustment	**Ignore**

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

Power	N1/N2	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation1 (SD1)	Standard Deviation2 (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.

Example4 - 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Higher is	**Good**
Test Type	**Non-Inferiority**
|E| (Equivalence Margin)	**10**
D (True Difference)	**0**
N1	*Ignored since this is the Find setting*
S (Std Deviation)	**40**
Alpha	**0.025**
Beta	**0.10**
Nonparametric Adjustment	**Ignore**
Population Size	**Infinite**

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

Power	N1/N2	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation1 (SD1)	Standard Deviation2 (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 Test of the Ratio of Two Means

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.
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

$$COV_Y = \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y} = \sqrt{e^{\sigma_X^2} - 1}$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Find

This option specifies the parameter to be solved for from the other parameters. In most situations, you will select either Beta for a power analysis or *N1* for sample size determination.

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.

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.

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.

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.

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.

Alpha (Significance Level)

Specify one or more values of alpha (the probably of a type-I error which is rejecting the null hypothesis of inferiority) when in fact the treatment group is not inferior to the reference group. Note that the valid range is 0 to 1, but typical values are between 0.01 and 0.20.

You can enter a range of values such as *0.05, 0.10, 0.15* or *0.5 to 0.15 by 0.01*.

Beta (1 - Power)

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
High Is	**Good**
Test Type	**Non-Inferiority**
E	**0.20**
R1	**0.95 1.0**
COV	**1.50**
Alpha	**0.025**
Beta	*Ignored since this is the Find setting*
N1	**100 to 1000 by 100**
N2	**Use R**
R	**1.0**

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)

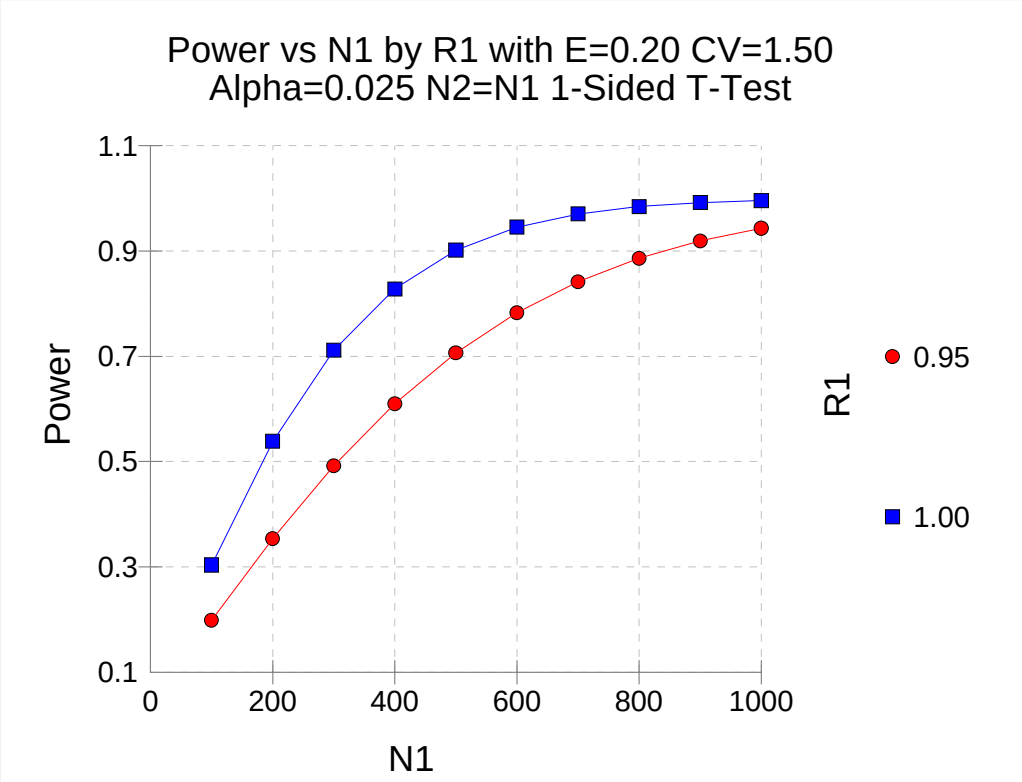
Power	N1/N2	Equivalence Margin (E)	Equivalence Bound (RB)	True Ratio (R1)	Significance Level (Alpha)	Coefficient of Variation (COV)	Beta
0.1987	100/100	0.20	0.80	0.95	0.0250	1.50	0.8013
0.3539	200/200	0.20	0.80	0.95	0.0250	1.50	0.6461
0.4918	300/300	0.20	0.80	0.95	0.0250	1.50	0.5082
0.6098	400/400	0.20	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.

Plot Section



This plot shows the power versus the sample size.

Example2 –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.

$$\begin{aligned} S1 &= \sqrt{\ln\left(COV^2 + 1\right)} \\ &= \sqrt{\ln\left(1.5^2 + 1\right)} \\ &= 1.085659 \\ E' &= \ln\left(1 - E\right) \\ &= \ln\left(0.8\right) \\ &= 0.223144 \\ D &= \ln\left(R1\right) \\ &= \ln\left(0.95\right) \\ &= -0.051293 \end{aligned}$$

Setup

Load the *PASS: Means: 2: Non-Inferiority: Differences* panel. You can enter the following parameter values or load Example1b.

Option	Value
Data Tab	
Find	**Beta and Power**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**0.223144**
D	**-0.051293 0.0**
S1	**1.085659**
S2	**S1**
Alpha	**0.025**
Beta	*Ignored since this is the Find setting*
N1	**100 to 1000 by 100**
N2	**Use R**
R	**1.0**

Annotated 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

Power	N1/N2	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation1 (SD1)	Standard Deviation2 (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 Example1 to validate the procedure. You will find that the power values are identical.

Chapter 460

Equivalence of Two Independent Means using their Difference

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 Labels or Plot Setup, turn to the chapter entitled *Procedure Templates*.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Find

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.

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

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

|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.

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.

S (Standard Deviation)

Specify the within-group standard deviation, σ. The standard deviation is assumed to be the same for both groups.

Alpha (Significance Level)

This option specifies one or more values for the significance level, alpha. 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 was used for alpha, but some statisticians recommend a value of 0.1 or even 0.2 in equivalence trials. An alpha of 0.05 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.

Beta (1 - Power)

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 nonequivalent means when in fact the means are equivalent.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Now the use of 0.10 is standard. You should pick a value for beta that represents the risk of a type-II error you are willing to take.

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.

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*.

Example1 - 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Upper Equivalence Limit	**19.2**
Lower Equivalence Limit	**-Upper Limit**
D	**-4**
Alpha	**0.05**
Beta	***Ignored***
S	**18**
N1	**3 5 8 10 15 20 30 40 50 60**
N2	**Use R**
R	**1.0**

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

Power	Reference Group Sample Size (N1)	Treatment Group Sample Size (N2)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard 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

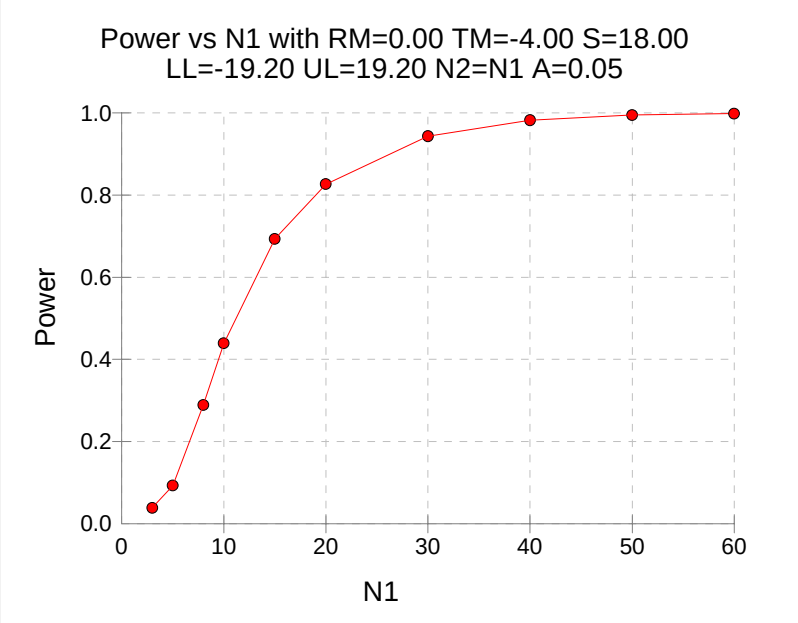
Report Definitions
Power is the probability of rejecting non-equivalence when they are equivalent.
N1 is the number of subjects in the reference group.
N2 is the number of subjects in the treatment group.
The Upper & Lower Limits are the maximum allowable differences that result in equivalence.
True 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.

Example2 - 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 ***PASS's***.

We will now setup this example in ***PASS***.

Setup

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N1**
D	**-2**
Upper Equivalence Limit	**5**
Lower Equivalence Limit	**-Upper Limit**
Alpha	**0.05**
Beta	**0.20**
S	**8**
N1	**Ignored**
N2	**Use R**
R	**1.0**

Annotated Output

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

Numeric Results

Power	Reference Group Sample Size (N1)	Treatment Group Sample Size (N2)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard 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 of Two Means using Simulation

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

$$H0\text{:}\left(\mu_1 - \mu_2\right) \le -D \text{ or } \left(\mu_1 - \mu_2\right) \ge D$$

where D is the margin of equivalence. Thus the null hypothesis is made up of two simple hypotheses:

$$H0_1\text{:}\left(\mu_1 - \mu_2\right) \le -D$$

$$H0_2\text{:}\left(\mu_1 - \mu_2\right) \ge D$$

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

$$s^*_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2}{N_1\left(N_1 - 1\right)}\right) + \left(\frac{\sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_2\left(N_2 - 1\right)}\right)}$$

$$f = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\left(N_1 - 1\right)} + \frac{s_2^4}{N_2^2\left(N_2 - 1\right)}}$$

$$s_1 = \sqrt{\left(\frac{\sum_{i=1}^{N_1}\left(X_{1i} - \overline{X}_1\right)^2}{N_1 - 1}\right)}, s_2 = \sqrt{\left(\frac{\sum_{i=1}^{N_2}\left(X_{2i} - \overline{X}_2\right)^2}{N_2 - 1}\right)}$$

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

$$\frac{1}{f} = \frac{c^2}{h_1 - 1} + \frac{1 - c^2}{h_2 - 1}$$

$$c = \frac{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)}}{\frac{SSD_{1wg}}{h_1\left(h_1 - 1\right)} + \frac{SSD_{2wg}}{h_2\left(h_2 - 1\right)}}$$

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 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.

Data Tab

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

Find

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.

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.

Equiv. 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.

Test Statistic

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).

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*.

Group 1 (and 2) Dist'n | 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) Dist'n | 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.

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). 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. Commonly, the value of 0.05 is used for equivalence tests.

Beta (1 - Power)

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 a false null hypothesis. Values must be between zero and one. Historically, the value of 0.20 was used for beta. Now, 0.10 is more common. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as one minus beta. Hence, specifying beta also specifies the power. 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.

Options Tab

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

Maximum Iterations

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 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.

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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N1	**10 30 50 70**
N2	**Use R**
R	**1.0**
Group 1 Dist'n | H0	**N(M0 S)**
Group 2 Dist'n | H0	**N(M1 S)**
Group 1 Dist'n | H1	**N(M0 S)**
Group 2 Dist'n | H1	**N(M0 S)**
M0 (Mean under H0)	**63**
M1 (Mean under H1)	**66**
S	**5**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*

Annotated Output

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

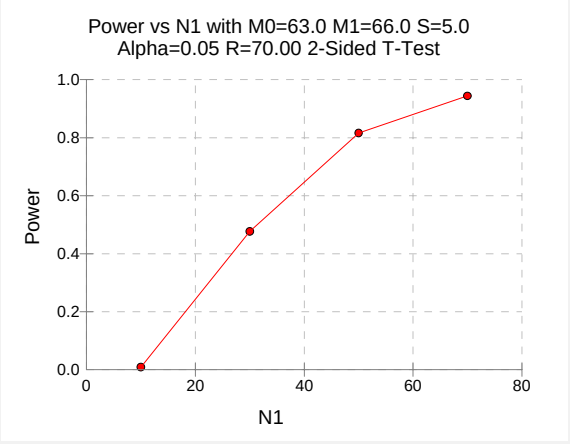
Numeric Results for Testing Mean Equivalence. Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S) & Normal(M1 S)
H1 Dist's: Normal(M0 S) & Normal(M0 S)
Test Statistic: T-Test

Power	N1/N2	H1 Diff1	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	Actual 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.

Example2 - 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N1**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
Group 1 Dist'n | H0	**N(M0 S)**
Group 2 Dist'n | H0	**N(M1 S)**
Group 1 Dist'n | H1	**N(M0 S)**
Group 2 Dist'n | H1	**N(M0 S)**
M0 (Mean under H0)	**63**
M1 (Mean under H1)	**66**
S	**5**
Alpha	**0.05**
Beta	**0.10**

Annotated Output

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

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

Power	N1/N2	H1 Diff1	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	Actual 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]		

Notes:
Pool Size: 10000. Simulations: 2000. Run Time: 49.81 seconds.

The required sample size is 61 per group.

Example3 – 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N1	**40**
N2	**Use R**
R	**1.0**
Group 1 Dist'n | H0	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Dist'n | H0	**N(M1 S)[95];N(M1 A)[5]**
Group 1 Dist'n | H1	**N(M0 S)[95];N(M0 A)[5]**
Group 2 Dist'n | H1	**N(M0 S)[95];N(M0 A)[5]**
M0 (Mean under H0)	**63**
M1 (Mean under H1)	**66**
S	**5**
A	**5 25 50**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

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

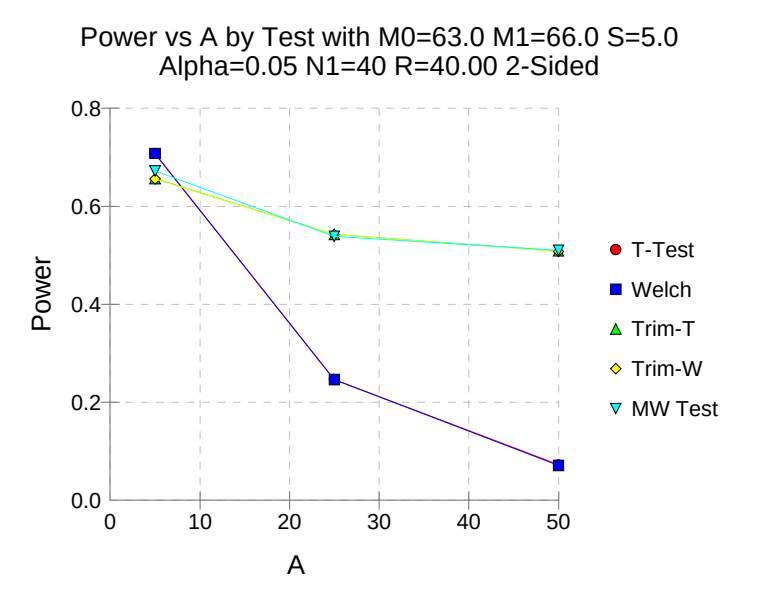
Power Comparison for Testing Equivalence. Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M1 S)[95];Normal(M1 A)[5]
H1 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M0 S)[95];Normal(M0 A)[5]

N1/N2	H1 Diff (Diff1)	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	T-Test Power	Welch Power	Trim. T-Test Power	Trim. Welch Power	Mann Whit'y 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|

N1/N2	H1 Diff (Diff1)	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	T-Test Alpha	Welch Alpha	Trim. T-Test Alpha	Trim. Welch Alpha	Mann Whit'y 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.

Example4 – 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N1	**40**
N2	**Use R**
R	**1.0**
Group 1 Dist'n | H0	**L(M0 S G 0)**
Group 2 Dist'n | H0	**L(M1 S G 0)**
Group 1 Dist'n | H1	**L(M0 S G 0)**
Group 2 Dist'n | H1	**L(M0 S G 0)**
M0 (Mean under H0)	**63**
M1 (Mean under H1)	**66**
S	**5**
G	**0 0.5 0.9**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

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

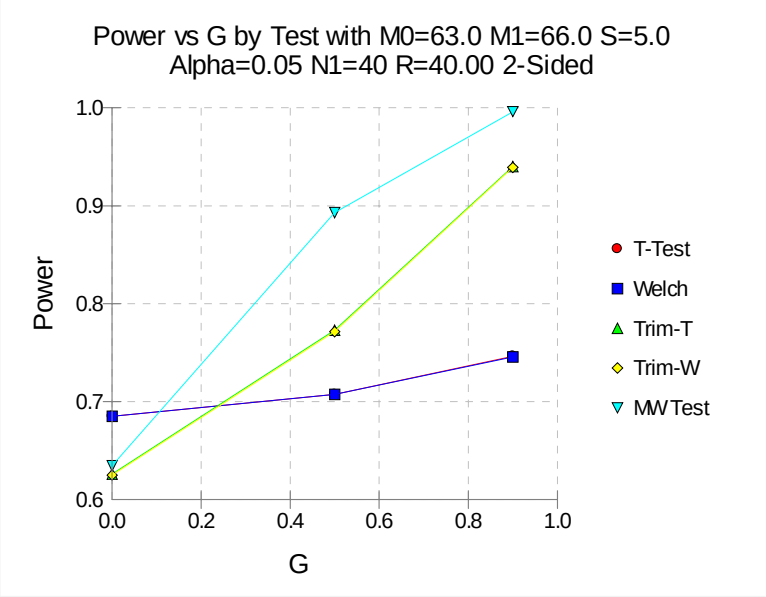
Power Comparison for Testing Equivalence. Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M1 S)[95];Normal(M1 A)[5]
H1 Dist's: Normal(M0 S)[95];Normal(M0 A)[5] & Normal(M0 S)[95];Normal(M0 A)[5]

N1/N2	H1 Diff (Diff1)	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	T-Test Power	Welch Power	Trim. T-Test Power	Trim. Welch Power	Mann Whit'y 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|

N1/N2	H1 Diff (Diff1)	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	T-Test Alpha	Welch Alpha	Trim. T-Test Alpha	Trim. Welch Alpha	Mann Whit'y 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.

Example5 – 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. 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	**N1**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1.0**
Group 1 Dist'n | H0	**N(M0 S)**
Group 2 Dist'n | H0	**N(91 S)**
Group 1 Dist'n | H1	**N(M0 S)**
Group 2 Dist'n | H1	**N(94 S)**
M0 (Mean under H0)	**96**
S	**8**
Alpha	**0.05**
Beta	**0.20**

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

Power	N1/N2	H1 Diff1	Lower Equiv. Limit	Upper Equiv. Limit	Target Alpha	Actual 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 of Two Independent Means using their Ratio

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 another chapter and they 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). They will only be summarized here.

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

$$\mathrm{H}_0: \phi \le \phi_L \text{ or } \phi \ge \phi_U \text{ where } \phi_L < 1, \phi_U > 1.$$

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.

$$\phi_L \le \phi \le \phi_U$$
$$\Rightarrow \phi_L \le \left\{\frac{\mu_T}{\mu_R}\right\} \le \phi_U$$
$$\Rightarrow \ln(\phi_L) \le \left\{\ln(\mu_T) - \ln(\mu_R)\right\} \le \ln(\phi_U)$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Find

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.

RU (Upper Equiv. 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 Equiv. 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.

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.

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 $\sigma_d^2 = 2\sigma_w^2$.

N1 (Sample Size Ref. 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 Trt. 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 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.

Alpha (Significance Level)

Specify one or more values of alpha, the probability of a type-I error which is rejecting the null hypothesis of non-equivalence when in fact the groups are equivalent. Note that the valid range is 0 to 1, but typical values are between 0.01 and 0.20.

You can enter a range of values such as *0.05, 0.10, 0.15* or *0.05 to 0.15 by 0.01*.

Beta (1 - Power)

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 non-equivalence when in fact the treatment mean is equivalent.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Currently, 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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
RU	**1/RL**
RL	**0.80**
R1	**1.0 1.05**
COV	**1.50**
N1	**50 to 550 by 100**
N2	**Use R**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*

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

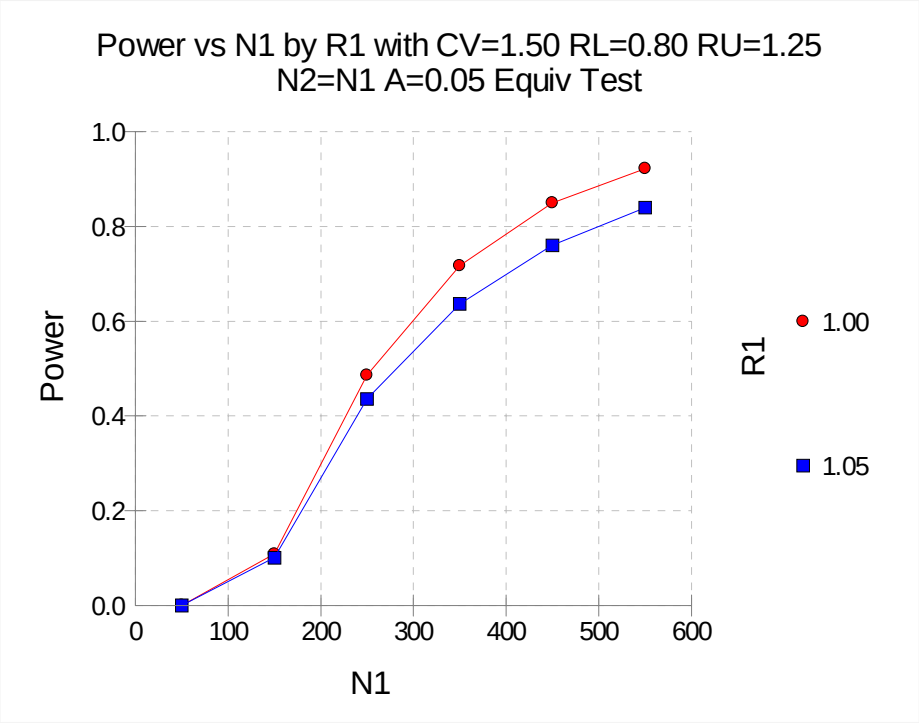
Power	Reference Group Sample Size (N1)	Treatment Group Sample Size (N2)	Lower Equiv. Limit (RL)	Upper Equiv. Limit (RU)	True Ratio (R1)	Coefficient of Variation (COV)	Alpha	Beta
0.0000	50	50	0.80	1.25	1.00	1.50	0.0500	1.0000
0.1088	150	150	0.80	1.25	1.00	1.50	0.0500	0.8912
0.4863	250	250	0.80	1.25	1.00	1.50	0.0500	0.5137
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.

Example2 –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

Load the panel. You can enter the following parameter values or load Example2.

Option	Value
Data Tab	
Find	**N1**
RU	**1/RL**
RL	**0.80**
R1	**1.0**
COV	**0.80**
N1	*Ignored since this is the Find setting*
N2	**Use R**
R	**1**
Alpha	**0.05**
Beta	**0.10**

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

Power	Reference Group Sample Size (N1)	Treatment Group Sample Size (N2)	Lower Equiv. Limit (RL)	Upper Equiv. Limit (RU)	True Ratio (R1)	Coefficient of Variation (COV)	Alpha	Beta
0.9012	216	216	0.80	1.25	1.00	0.80	0.0500	0.0988

PASS has also calculated the per group sample size to be 216 which matches Julious's result.

Chapter 475

Group Sequential Tests of 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:

$\alpha(0) = 0$

$\alpha(1) = \alpha$

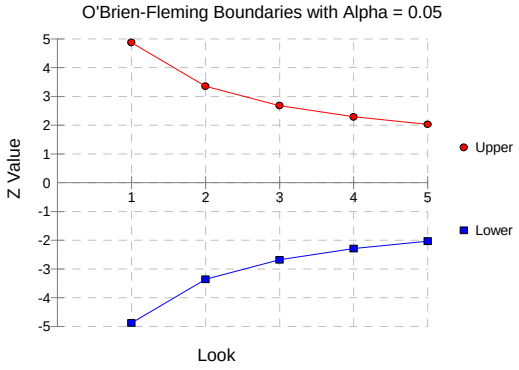
The last characteristic guarantees a fixed α level when the trial is complete. That is,

$$\Pr\left(|z_1| \geq b_1 \text{ or } |z_2| \geq b_2 \text{ or } \cdots \text{ or } |z_k| \geq b_k\right) = \alpha(\tau)$$

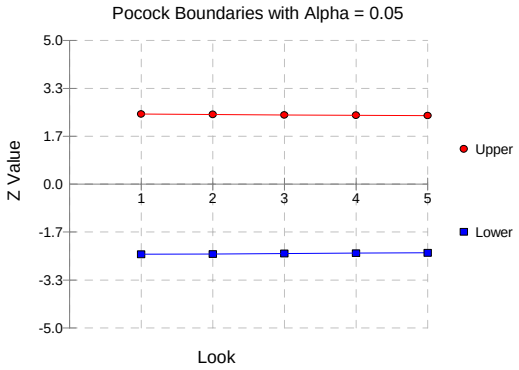
This methodology is very flexible since neither the times nor the number of analyses must be specified in advance. Only the functional form of $\alpha(\tau)$ must be specified.

PASS provides five popular spending functions plus the ability to enter and analyze your own boundaries. These are calculated as follows:

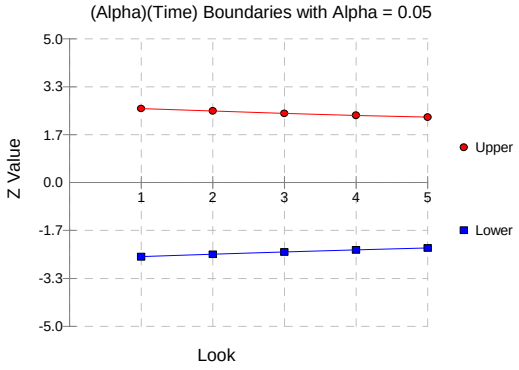
1. O'Brien-Fleming $2-2\Phi\left(\frac{Z_{\alpha/2}}{\sqrt{\tau}}\right)$



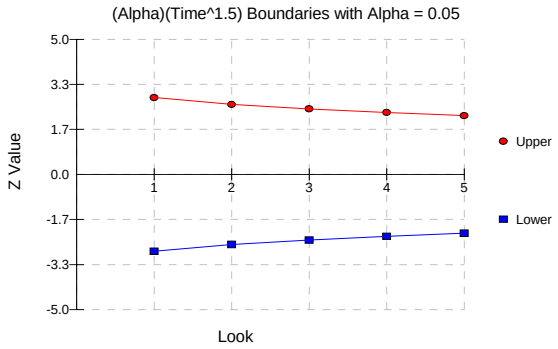
2. Pocock $\alpha\ln\left(1+\left(e-1\right)\tau\right)$



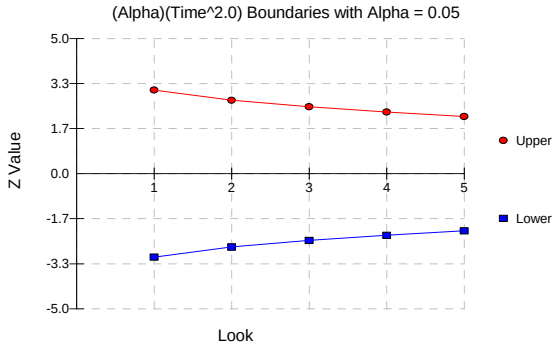
3. Alpha * time $\alpha\tau$



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, 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.

Data Tab

The Data tab contains the parameters associated with the *z* test such as the means, variances, sample sizes, alpha, and beta.

Find

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

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

Select *Beta* when you want to calculate the power of an experiment that has already been run since power is equal to one minus beta.

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).

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*.

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.

Alpha

This option specifies one or more values for the probability of a type-I error, alpha. This is also called the *significance level* or *test size*. A type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values of alpha must be between zero and one. Often, the value of 0.05 is used for alpha since this value is spread across several interim tests. This means that about one trial in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error that you are willing to take.

Beta (1-Power)

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 equal means when in fact they are different.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Now, 0.10 is more common. You should pick a value that represents the risk of a type-II error you are willing to take.

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.

S1 (Std Deviation 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 (Std Deviation 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*.

Sequential Tab

The Sequential tab 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.

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.

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.

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.

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.

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

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.

Options Tab

The Options tab controls the convergence of the various iterative algorithms used in the calculations.

Max Iterations 1

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).

Max Iterations 2

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.

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.

Bnd 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.

Example1 - 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**N1**
Mean1	**0**
Mean2	**10 to 70 by 10**
N1	**Ignored**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
Beta	**0.10**
S1	**44**
S2	**S1**
Sequential Tab	
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

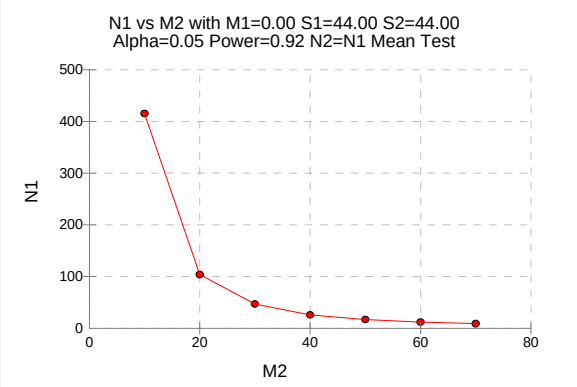
Report Definitions
Power is the probability of rejecting a false null hypothesis. Power should be close to one.
N1 and N2 are the number of items sampled from groups 1 and 2.
Alpha is the probability of rejecting a true null hypothesis in at least one of the sequential tests.
Beta is the probability 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

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total 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.

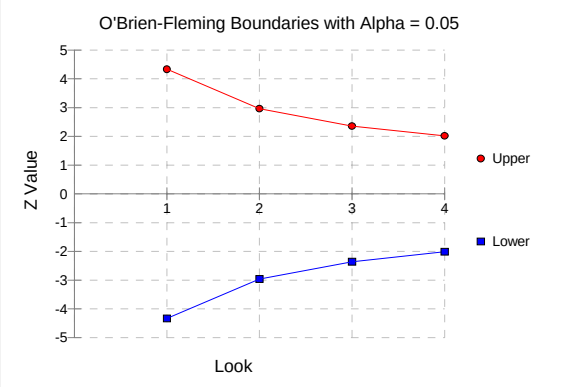
Total Power

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.

Example2 - 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean1	**0**
Mean2	**20**
N1	**20 to 220 by 40**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.01, 0.05, 0.10**
Beta	**Ignored**
S1	**44**
S2	**S1**
Sequential Tab	
Number of Looks	**4**
Spending Function	**O'Brien-Fleming**
Times	**Equally Spaced**
Max Time	**2**

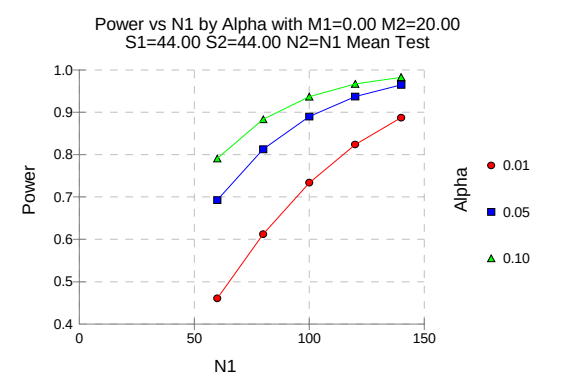
Annotated Output

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
0.1256	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.

Example3 - 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

You can enter these values yourself or load the Example3 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean1	**0**
Mean2	**20**
N1	**100**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
Beta	**Ignored**
S1	**44**
S2	**S1**
Sequential Tab	
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**

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.

Example4 - 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

You can enter these values yourself or load the Example4 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Mean1	**0**
Mean2	**20**
N1	**100**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05 (will be calculated from boundaries)**
Beta	**Ignored**
S1	**44**
S2	**S1**
Sequential Tab	
Number of Looks	**5**
Spending Function	**User Supplied**
Times	**Equally Spaced**
Lower Boundaries	**Symmetric**
Upper Boundaries	**3.5, 3.5, 3.0, 2.5, 2.0**
Max Time	**2**

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

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total 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.

Example5 - 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

You can enter these values yourself or load the Example5 template from the Template tab.

Option	Value
Data Tab	
Find	**N1**
Mean1	**220**
Mean2	**200**
N1	**Ignored**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
Beta	**0.10**
S1	**30**
S2	**S1**

Sequential Tab

Number of Looks.....**5**
Spending Function.....**O'Brien-Fleming**
Times.....**Equally Spaced**
Lower Boundaries.....**blank**
Upper Boundaries.....
Max Time.....**1**

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

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total 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.

Example6 - 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

You can enter these values yourself or load the Example6 template from the Template tab.

Option	Value
Data Tab	
Find	**N1**
Mean1	**220**
Mean2	**200**
N1	**Ignored**
N2	**Use R**
R	**1.0**
Alternative Hypothesis	**Two-Sided**
Alpha	**0.05**
Beta	**0.10**
S1	**30**
S2	**S1**
Sequential Tab	
Number of Looks	**5**
Spending Function	**Pocock**
Times	**Equally Spaced**
Lower Boundaries	**blank**
Upper Boundaries	
Max Time	**1**

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

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total 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

Two Means - Cluster Randomization

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

$$df = 2(M-1)$$

$$\lambda = \frac{d}{\left[2\left(1+\left(N-1\right)\rho\right)/\left(MN\right)\right]^{1/2}}$$

$$d = \frac{\left|\mu_1 - \mu_2\right|}{\sigma}$$

Procedure Options

This section describes the options that are unique to this procedure. These are located on the panels associated with the Data, 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.

Data Tab

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

Find

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 *Beta* (or *Power*).

Under most situations, you will select either *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 beta directly. For the other parameters, a search is made using an iterative procedure until an appropriate value is found.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject the null hypothesis when in fact it is true.

Values between 0.001 and 0.100 are acceptable. The value of 0.05 has become the standard. This means that about one test in twenty will falsely reject the null hypothesis. Although 0.05 is the standard value, you should pick a value for alpha that represents the risk of a type-I error you are willing to take.

Note that you can enter a range of values such as *0.01,0.05* or *0.01 to 0.05 by 0.01*.

Beta (1 - Power)

This option specifies one or more values for beta (the probability of accepting a false null hypothesis). Since statistical power is equal to one minus beta, specifying beta implicitly specifies the power. For example, setting beta at 0.20 also sets the power to 0.80.

Values must be between zero and one. The value of 0.20 has often used for beta. However, you should pick a value for beta that represents the risk of this type of error you are willing to take.

Note that you can enter a range of values such as *0.10,0.20* or *0.05 to 0.20 by 0.05*.

If your only interest is in determining the appropriate sample size for a confidence interval, set beta to 0.5.

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.

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.

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.

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.

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*.

Options Tab

This tab sets an option used in the iterative procedures.

Maximum Iterations

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.

Example1 - 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Alpha	**0.01, 0.05**
Beta	*Ignored since this is the Find parameter*
Alternative Hypothesis	**Two-Sided**
D	**0.2**
R	**0.01**
S	**1.0**
M	**6**
N	**50 to 300 by 50**

Annotated Output

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

Numeric Results

Numeric Results for Two-Sided Test

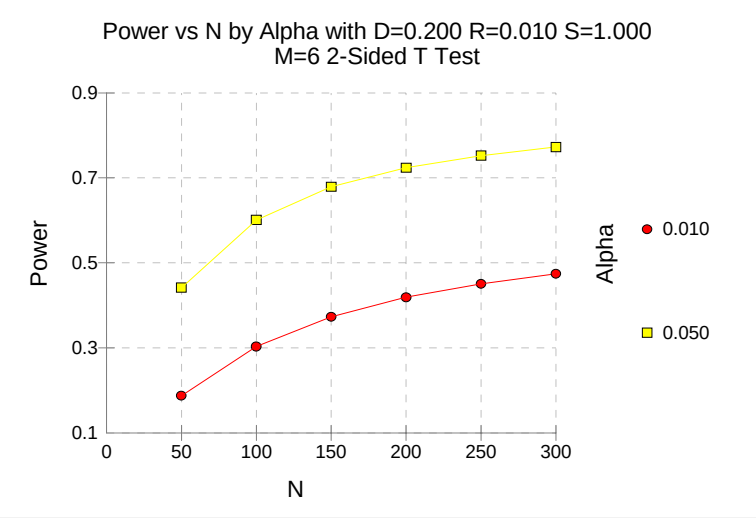
Power	M Number of Clusters	N Individuals Per Clusters	D Difference	R Intracluster Correlation	S Standard Deviation	Alpha	Beta
0.18754	6	50	0.200	0.01000	1.000	0.01000	0.81246
0.44200	6	50	0.200	0.01000	1.000	0.05000	0.55800
0.30320	6	100	0.200	0.01000	1.000	0.01000	0.69680
0.60128	6	100	0.200	0.01000	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.

Plot Section



This plot shows the power versus the cluster size for the two alpha values.

Example2 - 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Alpha	**0.05**
Beta	*Ignored since this is the Find parameter*
Alternative Hypothesis	**Two-Sided**
D	**0.2**
R	**0.001**
S	**1.0**
M	**3**
N	**100 300 500**

Annotated Output

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

Numeric Results

Numeric Results for Two-Sided Test

Power	M Number of Clusters	N Individuals Per Clusters	D Difference	R Intracluster Correlation	S Standard 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

Tests of Paired Means using Simulation

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 $M0$ 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 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.

Data Tab

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

Find

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.

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.

H1 (Alternative)

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**.

Test Statistic

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you falsely reject the null hypothesis. Values must be between zero and one. Commonly, the value of 0.05 is used for two-tailed tests and 0.025 is used for one-tailed tests.

Beta (1 - Power)

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 equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.20 was used for beta. Now, 0.10 is more common. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as one minus beta. Hence, specifying beta also specifies the power. 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.

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.

M0 (Item Mean Assuming 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 (Item Mean Assuming 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

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

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.

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.

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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N	**50 100 150**
R	**0.2**
Item A Distribution Assuming H0	**N(M0 S)**
Item B Distribution Assuming H0	**N(M0 S)**
Item A Distribution Assuming H1	**N(M0 S)**
Item B Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**
Alpha	**0.05**
Beta	*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 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

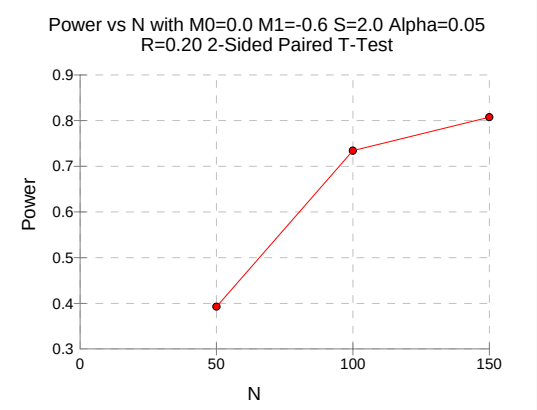
Power	N	H0 Diff0	H1 Diff1	Corr R	Target Alpha	Actual Alpha	Beta	M0	M1	S
0.393	50	0.0	-0.6	0.200	0.050	0.055	0.608	0.0	0.6	2.0
(0.021)	[0.371	0.414]				(0.010)	[0.045	0.064]		
0.734	100	0.0	-0.6	0.200	0.050	0.050	0.266	0.0	0.6	2.0
(0.019)	[0.715	0.753]				(0.010)	[0.040	0.060]		
0.808	150	0.0	-0.6	0.200	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.

Example2 - 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N	*Ignored since this is the Find setting*
R	**0.2**
Alpha	**0.05**
Beta	**0.10**
Item A Distribution Assuming H0	**N(M0 S)**
Item B Distribution Assuming H0	**N(M0 S)**
Item A Distribution Assuming H1	**N(M0 S)**
Item B Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**

Annotated Output

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

Numeric Results of Search for N

Power	N	H0 Diff0	H1 Diff1	Corr R	Target Alpha	Actual 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.

Example3 – Comparative results

Continuing with Example2, 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N	**50 100 150 200**
R	**0.2**
Item A Distribution Assuming H0	**N(M0 S)**
Item B Distribution Assuming H0	**N(M0 S)**
Item A Distribution Assuming H1	**N(M0 S)**
Item B Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

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

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)

N	H0 Diff (Diff0)	H1 Diff (Diff1)	Corr (R)	Target Alpha	T-Test Power	Wilcxn Power	Sign 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)

N	H0 Diff (Diff0)	H1 Diff (Diff1)	Corr (R)	Target Alpha	T-Test Alpha	Wilcxn Alpha	Sign 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.

Example4 - 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 Example1: 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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**10000**
H1 (Alternative)	**Diff<>Diff0**
Test Statistic	**T-Test**
N	**50**
R	**0.2**
Item A Distribution Assuming H0	**N(M0 S)**
Item B Distribution Assuming H0	**N(M0 S)**
Item A Distribution Assuming H1	**N(M0 S)**
Item B Distribution Assuming H1	**N(M1 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**0.6**
S	**2**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Options Tab	
Random Number Pool Size	**50000 (Increase to 5 times Simulations)**

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

Power	N	H0 Diff0	H1 Diff1	Corr R	Target Alpha	Actual 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.

Example5 – 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. 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	**Power**
Simulations	**2000**
H1 (Alternative)	**Diff>Diff0**
Test Statistic	**T-Test**
N	**5 10 15 20 25**
R	**0.2**
Item A Distribution Assuming H0	**N(M0 S)**
Item B Distribution Assuming H0	**N(M1 S)**
Item A Distribution Assuming H1	**N(M0 S)**
Item B Distribution Assuming H1	**N(M0 S)**
M0 (Mean under H0)	**0**
M1 (Mean under H1)	**5**
S	**5**
Alpha	**0.025**
Beta	*Ignored since this is the Find setting*

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

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

Power	N	H0 Diff0	H1 Diff1	Corr R	Target Alpha	Actual 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 of Paired Means Using Simulation

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

$$H0\colon (\mu_1 - \mu_2) \le -D \text{ or } (\mu_1 - \mu_2) \ge D$$

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 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, *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 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.

Data Tab

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

Find

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.

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.

Equiv. 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.

Test Statistic

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you falsely reject the null hypothesis. Values must be between zero and one. Commonly, the value of 0.05 is used for two-tailed tests and 0.025 is used for one-tailed tests.

Beta (1 - Power)

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 a false null hypothesis.

Values must be between zero and one. Historically, the value of 0.20 was used for beta. Now, 0.10 is more common. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as one minus beta. Hence, specifying beta also specifies the power. 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.

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.

M0 (Item A, or Reference, Mean)

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 (Item B, or Treatment, Mean)

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

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

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.

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.

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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N	**10 30 50 70**
R	**0.1 0.2**
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 (Item A Mean)	**63**
M1 (Item B Mean)	**66**
S	**5**
Alpha	**0.05**
Beta	*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 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

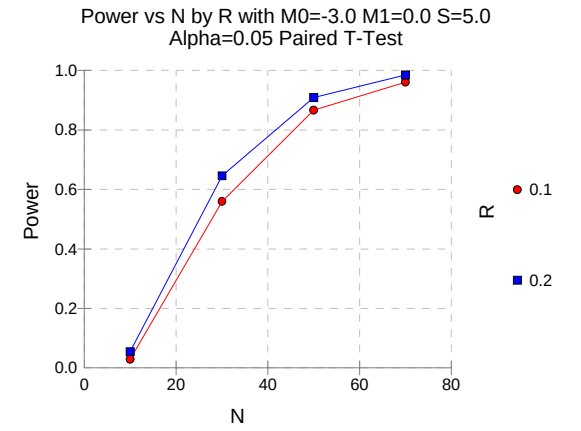
Power	N	H1 Diff1	Lower Equiv. Limit	Upper Equiv. Limit	Corr R	Target Alpha	Actual Alpha	M0	M1	S
0.030	10	0.0	-3.0	3.0	0.100	0.050	0.009	63.0	66.0	5.0
(0.007)	[0.022	0.037]					(0.004)	[0.005	0.013]	
0.055	10	0.0	-3.0	3.0	0.200	0.050	0.019	63.0	66.0	5.0
(0.010)	[0.045	0.065]					(0.006)	[0.013	0.025]	
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.

Example2 - 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? 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N	*Ignored since this is the Find setting*
R	**0.1 0.2**
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 (Item A Mean)	**63**
M1 (Item B Mean)	**66**
S	**5**
Alpha	**0.05**
Beta	**0.1**

Annotated Output

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

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

Power	N	H1 Diff1	Lower Equiv. Limit	Upper Equiv. Limit	Corr R	Target Alpha	Actual 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.

Example3 – Comparing Test Statistics

Continuing with Example2, 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N	**10 30 50 70**
R	**0.1**
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 (Item A Mean)	**63**
M1 (Item B Mean)	**66**
S	**5**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Reports Tab	
Show Comparative Reports	**Checked**
Show Comparative Plots	**Checked**

Annotated Output

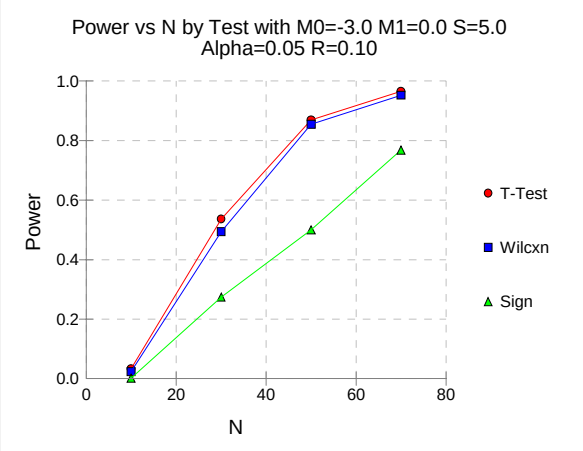
Click the Run button to perform the calculations and generate the following output.

Power Comparison for Testing Equivalence. Hypotheses: H0: Diff>=|Diff0|; H1: Diff<|Diff0|
H0 Dist'n: Normal(M0 S) - Normal(M1 S)
H1 Dist'n: Normal(M0 S) - Normal(M0 S)

N	H1 Diff (Diff1)	Lower Equiv. Limit	Upper Equiv. Limit	Corr (R)	Target Alpha	T-Test Power	Wilcxn Power	Sign 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)

N	H1 Diff (Diff1)	Lower Equiv. Limit	Upper Equiv. Limit	Corr (R)	Target Alpha	T-Test Alpha	Wilcxn Alpha	Sign 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.

Example4 - Validation

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. You can make these changes directly on your screen or you can load the template entitled Example4 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Power**
Simulations	**2000**
Equiv. Limit	**Symmetric**
Test Statistic	**T-Test**
N	**35 35 35 35 35**
R	**0.0**
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 (Item A Mean)	**0**
M1 (Item B Mean)	**0.05**
S	**0.070711**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*

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

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

Power	N	H1 Diff1	Lower Equiv. Limit	Upper Equiv. Limit	Corr R	Target Alpha	Actual 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 500

2x2 Cross-Over Designs for Comparing the Difference of Two Means

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

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

$$t_d = \frac{\left(\bar{x}_T - \bar{x}_R\right) - \delta}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

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: $\lambda = \frac{\delta\sqrt{N}}{\sigma_w\sqrt{2}}$.

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 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*.

Data Tab

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

Find

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

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

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

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*.

N (Sample Size of Both Groups)

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*.

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.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (false positive). A type-I error occurs when you reject the null hypothesis when in fact it is true.

Usually the value of 0.05 is used for a two-sided test and 0.025 for a one-sided test.

Beta (1 - Power)

This option specifies one or more values for the probability of a type-II error (false negative). A type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different. You cannot make both a type-I and a type-II error in a single hypothesis test.

Values must be between zero and one. Usually, the values of 0.10 or 0.20 are used for beta. However, you should pick a value for beta that represents the risk of a type-II error you are willing to take.

Power is defined as one minus beta. 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.

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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta and Power**
Diff0	**0**
Diff1	**5 10**
Specify S as Sw or Sd	**Sd**
S	**10**
Alternative Hypothesis	**H1: Diff0 <> Diff1**
N	**5 10 15 20 30 40 50**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*

Annotated Output

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

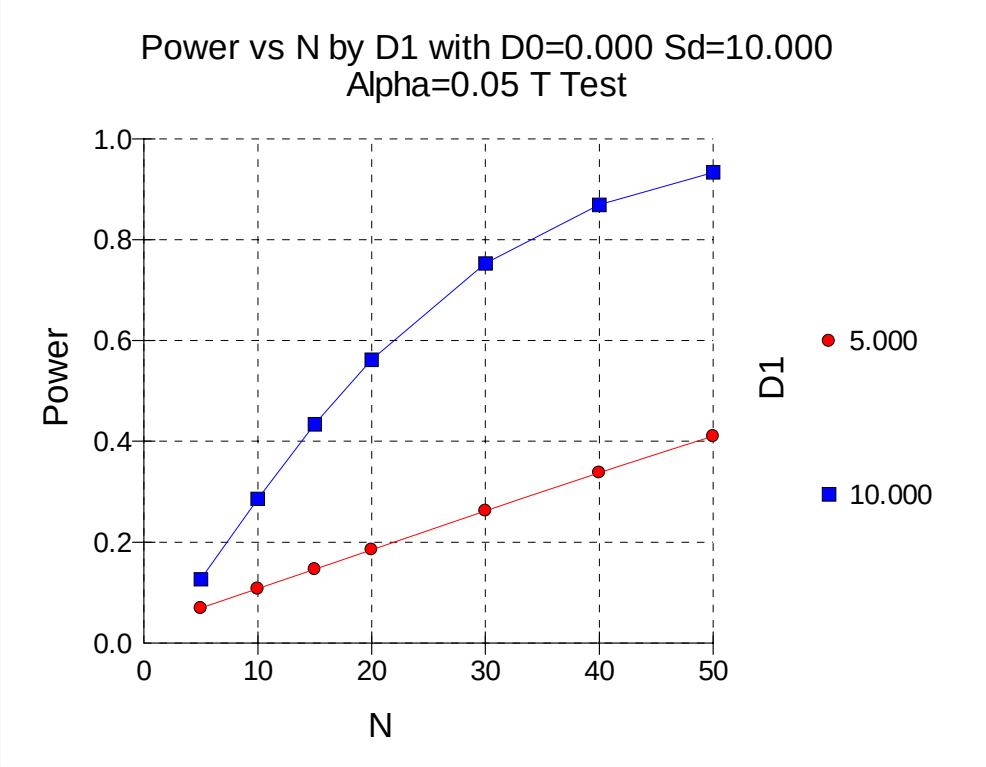
Numeric Results for Cross-Over Design
Null Hypothesis: Diff0=Diff1 Alternative Hypothesis: Diff0<>Diff1

Power	N	Diff0	Diff1	Alpha	Beta	Sd	Effect 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.

Chart Section



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.

Example2 - Finding the Sample Size

Continuing with Example1, 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. You can make these changes directly on your screen or you can load the template entitled Example2 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N (Sample Size)**
Diff0	**0**
Diff1	**5 10**
Specify S as Sw or Sd	**Sd**
Sd	**10**
Alternative Hypothesis	**H1: Diff0 <> Diff1**
Alpha	**0.05**
Beta	**0.10**
N	*Ignored since this is the Find setting*

Annotated Output

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

Numeric Results for Cross-Over Design
Null Hypothesis: Diff0=Diff1 Alternative Hypothesis: Diff0<>Diff1

Power	N	Diff0	Diff1	Alpha	Beta	Sd	Effect 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.

Example3 - 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N (Sample Size)**
Diff0	**0**
Diff1	**10**
Specify S as Sw or Sd	**Sw**
S	**20**
Alternative Hypothesis	**H1: Diff0 <> Diff1**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.10**

Annotated Output

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

Power	N	Diff0	Diff1	Alpha	Beta	Sw	Effect 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

2x2 Cross-Over Designs: Tests 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.

$$\phi = \phi_0$$
$$\Rightarrow \phi = \left\{\frac{\mu_T}{\mu_R}\right\}$$
$$\Rightarrow \ln(\phi) \neq \left\{\ln(\mu_T) - \ln(\mu_R)\right\}$$

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

$$COV_Y = \frac{\sqrt{\mu_Y^2 \left(e^{\sigma_X^2} - 1 \right)}}{\mu_Y}$$
$$= \sqrt{e^{\sigma_X^2} - 1}$$
$$= \sqrt{e^{\sigma_w^2} - 1}$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Find

This option specifies the parameter to be solved for from the other parameters. In most situations, you will select either Beta for a power analysis or *N1* for sample size determination.

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*.

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.

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 $\sigma_d^2 = 2\sigma_w^2$.

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*.

Alpha (Significance Level)

Specify one or more values of alpha, the probably of a type-I error which is rejecting the null hypothesis of equality when in fact the groups are different. Note that the valid range is 0 to 1, but typical values are between 0.01 and 0.20.

You can enter a range of values such as *0.05 0.10 0.15* or *0.5 to 0.15 by 0.05*.

Beta (1 - Power)

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 equal means when in fact the means are different.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Alternative Hypothesis	**R1<>R0 (Two-Sided)**
R0	**1.0**
R1	**1.1 1.2**
COV	**0.50**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
N	**20 to 220 by 40**

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.

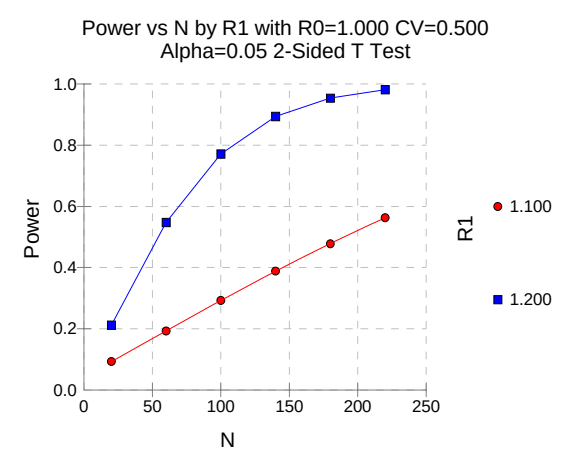
Power	Total Sample Size (N)	Mean Ratio Under H0 (R0)	Mean Ratio Under H1 (R1)	Effect Size (ES)	Coefficient of Variation (COV)	Significance Level (Alpha)	Beta
0.0928	20	1.000	1.100	0.143	0.500	0.0500	0.9072
0.1925	60	1.000	1.100	0.143	0.500	0.0500	0.8075
0.2925	100	1.000	1.100	0.143	0.500	0.0500	0.7075
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
0.9813	220	1.000	1.200	0.273	0.500	0.0500	0.0187

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.

Plot Section



This plot shows the power versus the sample size.

Example2 –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.

$$\begin{aligned} Sw &= \sqrt{\ln\left(COV^2 + 1\right)} \\ &= \sqrt{\ln\left(0.5^2 + 1\right)} \\ &= 0.472381 \end{aligned}$$

$$\begin{aligned} Diff1 &= \ln\left(R1\right) & Diff1 &= \ln\left(R1\right) \\ &= \ln\left(1.10\right) & &= \ln\left(1.20\right) \\ &= 0.095310 & &= 0.182322 \end{aligned}$$

Setup

Load the *PASS: Means: Two: Cross-Over: 2x2: Inequality: Differences* panel. You can enter the following parameter values or load Example1c.

Option	Value
Data Tab	
Find	**Beta and Power**
Diff0	**0**
Diff1	**0.095310 0.182322**
N	**20 to 220 by 40**
Alternative Hypothesis	**H1: DIFF0<>Diff1**
Specify S as Sw or Sd	**Sw**
Sw	**0.472381**
Alpha	**0.05**
Beta	*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 2x2 Cross-Over Design
Null Hypothesis: Diff0=Diff1 Alternative Hypothesis: Diff0<>Diff1

Power	N	Diff0	Diff1	Alpha	Beta	Sw	Effect 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
0.9813	220	0.000	0.182	0.0500	0.0187	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

2x2 Cross-Over Design: Non-Inferiority Tests 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

$$\mathrm{H}_0\colon \mu_X \leq A \text{ versus } \mathrm{H}_1\colon \mu_X > A$$

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*.

$$\mathrm{H}_0\colon \mu_X \geq A \text{ versus } \mathrm{H}_1\colon \mu_X < A$$

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 $-|\varepsilon|$.

$H_0: \mu_T \geq \mu_R - |\varepsilon|$ versus $H_1: \mu_T < \mu_R - |\varepsilon|$

$H_0: \mu_T - \mu_R \geq -|\varepsilon|$ versus $H_1: \mu_T - \mu_R < -|\varepsilon|$

$H_0: \delta \geq -|\varepsilon|$ versus $H_1: \delta < -|\varepsilon|$

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

$$t_d = \frac{\left(\bar{x}_T - \bar{x}_R\right) - \varepsilon}{\hat{\sigma}_w \sqrt{\frac{2}{N}}}$$

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 = $1 - T'_{df,\lambda}(t_\alpha)$, 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*.

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.

Data Tab

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

Find

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

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

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

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.

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.

|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.

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha). A type-I error occurs when you reject the null hypothesis of inferiority when in fact the treatment mean is non-inferior. Since this is a one-sided test, the value of 0.025 is commonly used for alpha.

Beta (1 - Power)

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 a false null hypothesis. You cannot make both a type-I and a type-II error in a single hypothesis test. Values must be between zero and one. The value of 0.10 is recommended for beta.

Power is defined as one minus beta. 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.

Example1 - 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. You can make these changes directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**Beta and Power**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**5 10**
D	**0**
Specify S as Sw or Sd	**Sw**
S	**10**
N	**5 10 15 20 30 40 50**
Alpha	**0.025**
Beta	*Ignored since this is the Find setting*

Annotated Output

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

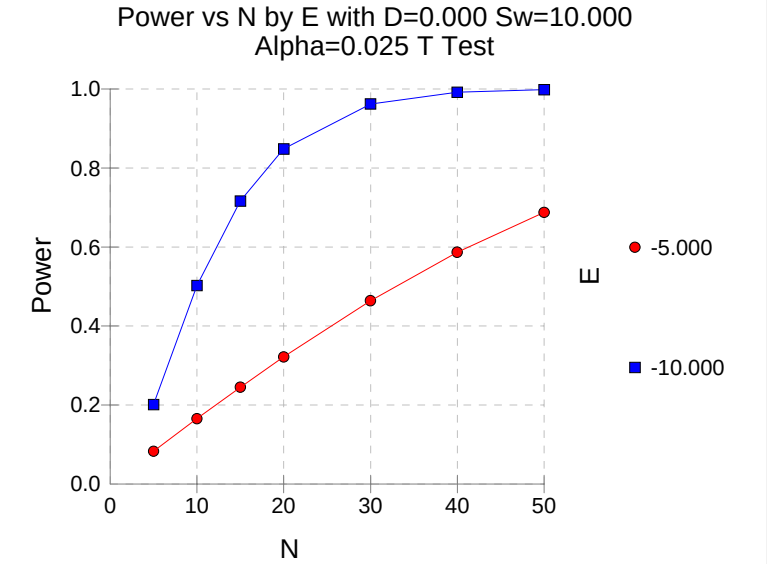
Numeric Results for Non-Inferiority T-Test (H0: D <= -|E|; H1: D > -|E|)

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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.

Chart Section



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.

Example2 - Finding the Sample Size

Continuing with Example1, suppose the researchers want to find the exact sample size necessary to achieve 90% power for both values of D.

You can enter the options below or load Example2 from the Template tab.

Option	Value
Data Tab	
Find	**N (Sample Size)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**5 10**
D	**0**
Specify S as Sw or Sd	**Sw**
S	**10**
Alpha	**0.025**
Beta	**0.10**
N	*Ignored since this is the Find setting*

Annotated Output

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

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

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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.

Example3 - 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. You can make these changes directly on your screen or you can load the template entitled Example3 by clicking the Template tab and loading this template.

Option	Value
Data Tab	
Find	**N (Sample Size)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**10**
D	**0**
Specify S as Sw or Sd	**Sw**
S	**20**
N	*Ignored since this is the Find setting*
Alpha	**0.025**
Beta	**0.10**

Annotated Output

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

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

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (Sw)
0.90657	88	-10.000	0.000	0.02500	0.09343	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

2x2 Cross-Over Designs: Non-Inferiority Tests 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

$$\mathrm{H}_0: \phi \le \phi_L \quad \textit{where } \phi_L < 1.$$

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.

$$\phi_L \le \phi$$
$$\Rightarrow \phi_L \le \left\{\frac{\mu_T}{\mu_R}\right\}$$
$$\Rightarrow \ln(\phi_L) \le \left\{\ln(\mu_T) - \ln(\mu_R)\right\}$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Find

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 *N* for sample size determination.

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.

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.

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.

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 $\sigma_d^2 = 2\sigma_w^2$.

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*.

Alpha (Significance Level)

Specify one or more values of alpha (the probably of a type-I error which is rejecting the null hypothesis of inferiority) when in fact the treatment group is not inferior to the reference group. Note that the valid range is 0 to 1, but typical values are between 0.01 and 0.20.

You can enter a range of values such as *0.05, 0.10, 0.15* or *0.5 to 0.15 by 0.01*.

Beta (1 - Power)

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
High Is	**Good**
Test Type	**Non-Inferiority**
E	**0.20**
R1	**1.0**
COV	**1.50**
N	**50 to 550 by 100**
Alpha	**0.05**
Beta	*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)

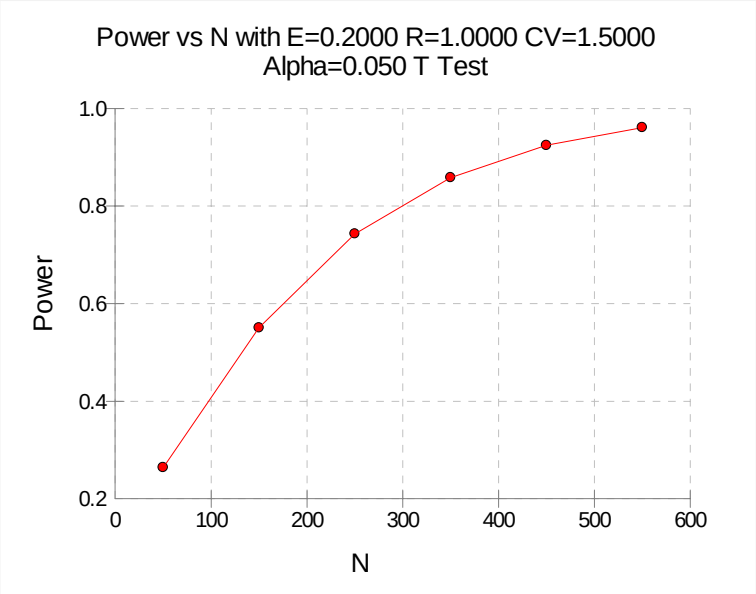
Power	N	Relative Equivalence Margin (E)	Ratio Equivalence Bound (RB)	Actual Ratio (R1)	Significance Level (Alpha)	Beta	COV
0.2638	50	0.2000	0.8000	1.0000	0.0500	0.7362	1.5000
0.5505	150	0.2000	0.8000	1.0000	0.0500	0.4495	1.5000
0.7431	250	0.2000	0.8000	1.0000	0.0500	0.2569	1.5000
0.8584	350	0.2000	0.8000	1.0000	0.0500	0.1416	1.5000
0.9246	450	0.2000	0.8000	1.0000	0.0500	0.0754	1.5000
0.9610	550	0.2000	0.8000	1.0000	0.0500	0.0390	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.

Example2 –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.

$$\begin{aligned} Sw &= \sqrt{\ln\left(COV^2 + 1\right)} \\ &= \sqrt{\ln\left(1.5^2 + 1\right)} \\ &= 1.085659 \\ E &= \sqrt{\ln\left(1 - E\right)} \\ &= \sqrt{\ln\left(0.8\right)} \\ &= 0.223144 \end{aligned}$$

Setup

Load the *PASS: Means: 2x2 Cross-Over: Non-Inferiority: Differences* panel. You can enter the following parameter values or load Example1a.

Option	Value
Data Tab	
Find	**Beta and Power**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**0.223144**
D	**0**
Specify S as Sw or Sd	**Sw**
S	**1.085659**
N	**50 to 550 by 100**
Alpha	**0.05**
Beta	*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 T-Test (H0: D <= -|E|; H1: D > -|E|)

Power	N	Equivalence Margin (E)	Actual Difference (D)	Significance Level (Alpha)	Beta	Standard Deviation (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
0.7431	250	-0.223	0.000	0.0500	0.2569	1.086
0.8584	350	-0.223	0.000	0.0500	0.1416	1.086
0.9246	450	-0.223	0.000	0.0500	0.0754	1.086
0.9610	550	-0.223	0.000	0.0500	0.0390	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

2x2 Cross-Over Design: Testing Equivalence 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

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

$$\mathrm{H}_0: \delta \le \varepsilon_L \text{ or } \delta \ge \varepsilon_U \text{ where } \varepsilon_L < 0, \varepsilon_U > 0.$$

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

$$\Pr(T_L \geq t_{1-\alpha,N-2} \text{ and } T_U \leq -t_{1-\alpha,N-2})$$

where T_L and T_U are distributed as the bivariate, noncentral *t* distribution with noncentrality parameters Δ_L and Δ_U given by

$$\Delta_L = \frac{\delta - \varepsilon_L}{\sigma_w \sqrt{\frac{2}{N}}}$$

$$\Delta_U = \frac{\delta - \varepsilon_U}{\sigma_w \sqrt{\frac{2}{N}}}$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Find

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

|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.

D (True Value)

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the significance level, alpha. 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 was used for alpha, but some statisticians recommend a value of 0.1 or even 0.2 in equivalence trials. An alpha of 0.05 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.

Beta (1 - Power)

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 nonequivalent means when in fact the means are equivalent.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Now the use of 0.10 is standard. You should pick a value for beta that represents the risk of a type-II error you are willing to take.

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.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
|EU|	**19.2**
-|EL|	**-Upper Limit**
D	**-4**
N	**6 10 16 20 40 60 80 100**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Specify S as Sw or Sd	**Sw**
S	**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

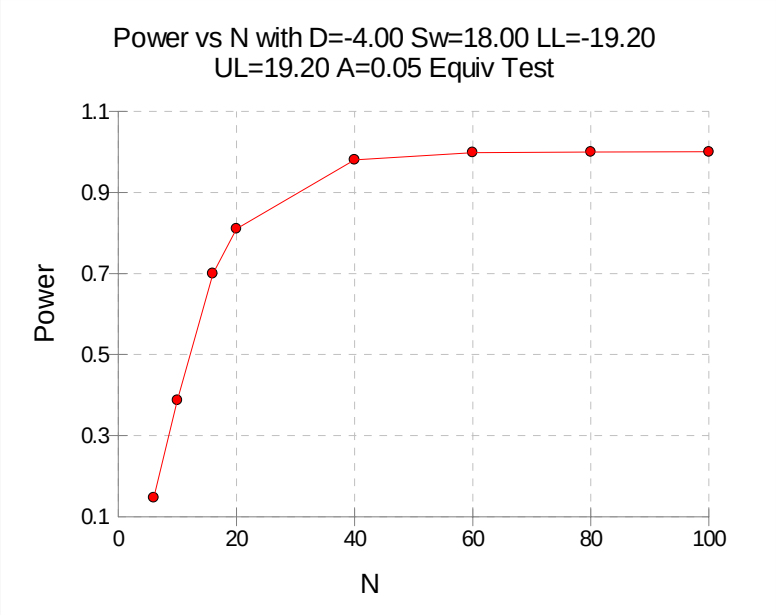
Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation Sw	Alpha	Beta
0.1470	6	-19.20	19.20	-4.00	18.00	0.0500	0.8530
0.3873	10	-19.20	19.20	-4.00	18.00	0.0500	0.6127
0.6997	16	-19.20	19.20	-4.00	18.00	0.0500	0.3003
0.8104	20	-19.20	19.20	-4.00	18.00	0.0500	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.

Plot Section



This plot shows the power versus the sample size.

Example2 – 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N**
|EU|	**19.2**
-|EL|	**-Upper Limit**
D	**-4**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.1 0.2**
Specify S as Sw or Sd	**Sw**
S	**18**

Annotated Output

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

Numeric Results for Testing Equivalence Using a Cross-Over Design

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation 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.

Example3 – 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

You can enter these values yourself or load the Example3 template from the Template tab.

Option	Value
Data Tab	
Find	**N**
|EU|	**20**
-|EL|	**-Upper Limit**
D	**0 -5 -10 -15**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.3**
Specify S as Sw or Sd	**Sw**
S	**20**

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

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation 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).

Example4 –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

You can enter these values yourself or load the Example4 template from the Template tab.

Option	Value
Data Tab	
Find	**N**
|EU|	**20**
-|EL|	**-Upper Limit**
D	**0**
N	*Ignored since this is the Find setting*
Alpha	**0.10**
Beta	**0.2**
Specify S as Sw or Sd	**Sw**
S	**40**

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

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation 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.

Example5 –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

Option	Value
Data Tab	
Find	**Beta and Power**
|EU|	**20**
-|EL|	**-Upper Limit**
D	**0**
N	**10 12 13 14 16**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
Specify S as Sw or Sd	**Sw**
S	**15.66**

You can enter these values yourself or load the Example5 template from the Template tab.

Annotated Output

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

Numeric Results for Testing Equivalence Using a Cross-Over Design

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation 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).

Example6 –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

You can enter these values yourself or load the Example6 template from the Template tab.

Option	Value
Data Tab	
Find	**N**
|EU|	**30**
-|EL|	**-Upper Limit**
D	**0**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.2**
Specify S as Sw or Sd	**Sw**
S	**45**

Annotated Output

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

Numeric Results for Testing Equivalence Using a Cross-Over Design

Power	Total Sample Size (N)	Lower Equiv. Limit	Upper Equiv. Limit	True Difference	Standard Deviation 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

2x2 Cross-Over Design: Testing Equivalence 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

$$H_0: \phi \le \phi_L \text{ or } \phi \ge \phi_U \text{ where } \phi_L < 1, \phi_U > 1.$$

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.

$$\phi_L \le \phi \le \phi_U$$
$$\Rightarrow \phi_L \le \left\{\frac{\mu_T}{\mu_R}\right\} \le \phi_U$$
$$\Rightarrow \ln(\phi_L) \le \left\{\ln(\mu_T) - \ln(\mu_R)\right\} \le \ln(\phi_U)$$

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

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and power.

Find

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 *N* for sample size determination.

RU (Upper Equiv. 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 Equiv. 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.

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.

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*.

Alpha (Significance Level)

Specify one or more values of alpha, the probability of a type-I error which is rejecting the null hypothesis of non-equivalence when in fact the groups are equivalent. Note that the valid range is 0 to 1, but typical values are between 0.01 and 0.20.

You can enter a range of values such as *0.05, 0.10, 0.15* or *0.05 to 0.15 by 0.01*.

Beta (1 - Power)

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 non-equivalence when in fact the treatment mean is equivalent.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta. Currently, 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.

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 $\sigma_d^2 = 2\sigma_w^2$.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
RU	**1/RL**
RL	**0.90**
R1	**1.0**
N	**50 to 550 by 100**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
COV	**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

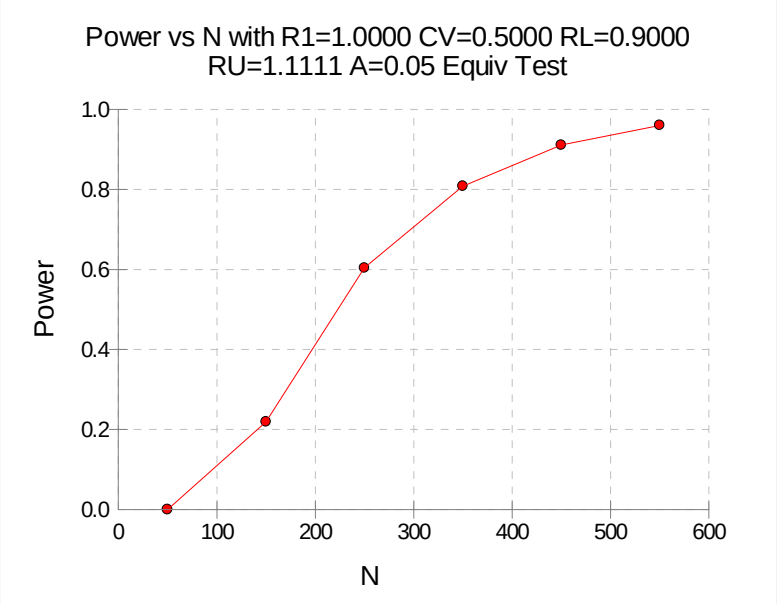
Power	Total Sample Size (N)	Lower Equiv. Limit of Ratio (RL)	Upper Equiv. Limit of Ratio (RU)	True Ratio (R1)	Coefficient of Variation (COV)	Alpha	Beta
0.0000	50	0.9000	1.1111	1.0000	0.5000	0.0500	1.0000
0.2190	150	0.9000	1.1111	1.0000	0.5000	0.0500	0.7810
0.6037	250	0.9000	1.1111	1.0000	0.5000	0.0500	0.3963
0.8079	350	0.9000	1.1111	1.0000	0.5000	0.0500	0.1921
0.9107	450	0.9000	1.1111	1.0000	0.5000	0.0500	0.0893
0.9598	550	0.9000	1.1111	1.0000	0.5000	0.0500	0.0402

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.

Plot Section



This plot shows the power versus the sample size.

Example2 –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.

We will now setup this example in ***PASS***.

Setup

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N**
RU	**1/RL**
RL	**0.90 0.85 0.80 0.75**
R1	**1.00**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.90**
COV	**0.25**

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

Power	Total Sample Size (N)	Lower Equiv. Limit of Ratio (RL)	Upper Equiv. Limit of Ratio (RU)	True Ratio (R1)	Coefficient of Variation (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

Higher-Order Cross-Over Designs: Non-Inferiority Tests 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

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

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 Non-Inferiority Test

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

$$H_0: \delta \leq A \text{ versus } H_1: \delta > A$$

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*.

$$H_0: \delta \geq A \text{ versus } H_1: \delta < A$$

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

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

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

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

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 specified value of δ must be greater than the specified value of $|\varepsilon|$. 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 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

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

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

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

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.

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Find

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

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.

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.

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.

|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 Value)

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|.

Sw (Within Std. 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.

Alpha (Significance Level)

This option specifies one or more values for the significance level, alpha. A type-I error occurs when you reject the null hypothesis when it is true.

Values must be between zero and one. The value of 0.05 is often used for alpha. An alpha of 0.05 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.

Beta (1 - Power)

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 a false null hypothesis.

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.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Design Type	**2x3 (Two-Sequence Dual)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**5 10**
D	**0**
N	**6 to 66 by 10**
Sw	**10**
Alpha	**0.025**
Beta	*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 Testing Non-Inferiority Using the Difference
Design: Two-Sequence Dual Cross-Over. Hypotheses: H0: D <= -\|E\|; H1: D > -\|E\|.

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Equivalence Margin |E|	Difference for Power (D)	Standard Error of Diff. (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
0.3837	6	2x3	10.00	0.00	10.00	0.0250	0.6163
0.8832	16	2x3	10.00	0.00	10.00	0.0250	0.1168
0.9818	26	2x3	10.00	0.00	10.00	0.0250	0.0182
0.9975	36	2x3	10.00	0.00	10.00	0.0250	0.0025
0.9997	46	2x3	10.00	0.00	10.00	0.0250	0.0003
1.0000	56	2x3	10.00	0.00	10.00	0.0250	0.0000
1.0000	66	2x3	10.00	0.00	10.00	0.0250	0.0000

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.

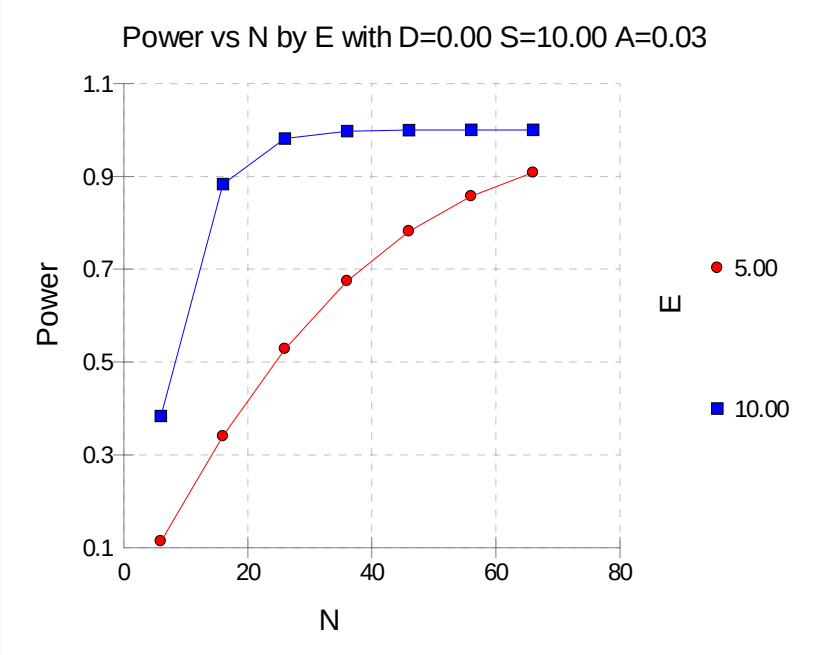
Chen, K.W.; Chow, S.C.; and Li, G. 1997. 'A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs.' Journal of Pharmacokinetics and Biopharmaceutics, Volume 25, No. 6, pages 753-765.

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.

Plot Section



This plot shows the power versus the sample size.

Example2 – 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

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**2x3 (Two-Sequence Dual)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**5 10**
D	**0**
N	*Ignored since this is the Find setting*
Sw	**10**
Alpha	**0.025**
Beta	**0.10 0.20**

You can enter these values yourself or load the Example2 template from the Template tab.

Annotated Output

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

Numeric Results for Testing Non-Inferiority Using the Difference
Design: Two-Sequence Dual Cross-Over. Hypotheses: H0: D <= -|E|; H1: D > -|E|.

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Equivalence Margin |E|	Difference for Power (D)	Standard Error of Diff. (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.

Example3 –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 High-Order, Cross-Over Equivalence 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. ***PASS*** calculates a sample size of 16. We will now setup this example in ***PASS***.

Setup

You can enter these values yourself or load the Example3 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**4x2 (Balaam)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
|E|	**0.2**
D	**0.05**
N	*Ignored since this is the Find setting*
Sw	**0.10**
Alpha	**0.05**
Beta	**0.10**

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: Balaam's Cross-Over. Hypotheses: H0: D <= -\|E\|; H1: D > -\|E\|.

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Equivalence Margin |E|	Difference for Power (D)	Standard Error of Diff. (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.

Chapter 535

Higher-Order Cross-Over Designs: Non-Inferiority Tests 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

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

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 Non-Inferiority Test

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

$$H_0: \phi \leq A \text{ versus } H_1: \phi > A$$

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*.

$$H_0: \phi \geq A \text{ versus } H_1: \phi < A$$

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

$$H_0: \phi \leq 1 - \varepsilon \text{ where } \varepsilon > 0.$$

The alternative hypothesis of non-inferiority is

$$H_1: \phi > 1 - \varepsilon$$

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.

$$1-\varepsilon < \phi$$
$$\Rightarrow 1-\varepsilon < \left\{\frac{\mu_T}{\mu_R}\right\}$$
$$\Rightarrow \ln(1-\varepsilon) < \left\{\ln(\mu_T) - \ln(\mu_R)\right\}$$

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

$$\mathrm{H}_0: \frac{\mu_T}{\mu_R} \le (1-\varepsilon) \quad \text{versus} \quad \mathrm{H}_1: \frac{\mu_T}{\mu_R} > (1-\varepsilon)$$

$$\mathrm{H}_0: \ln(\mu_T) - \ln(\mu_R) \le \ln(1-\varepsilon) \quad \text{versus} \quad \mathrm{H}_1: \ln(\mu_T) - \ln(\mu_R) > \ln(1-\varepsilon)$$

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

$$\mathrm{H}_0: \frac{\mu_T}{\mu_R} \geq (1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1: \frac{\mu_T}{\mu_R} < (1+\varepsilon)$$

$$\mathrm{H}_0: \ln(\mu_T) - \ln(\mu_R) \geq \ln(1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1: \ln(\mu_T) - \ln(\mu_R) < \ln(1+\varepsilon)$$

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

$$\mathrm{H}_0: \frac{\mu_T}{\mu_R} \leq (1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1: \frac{\mu_T}{\mu_R} > (1+\varepsilon)$$

$$\mathrm{H}_0: \ln(\mu_T) - \ln(\mu_R) \leq \ln(1+\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1: \ln(\mu_T) - \ln(\mu_R) > \ln(1+\varepsilon)$$

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

$$\mathrm{H}_0: \frac{\mu_T}{\mu_R} \geq (1-\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1: \frac{\mu_T}{\mu_R} < (1-\varepsilon)$$

$$\mathrm{H}_0: \ln(\mu_T) - \ln(\mu_R) \geq \ln(1-\varepsilon) \qquad \text{versus} \qquad \mathrm{H}_1: \ln(\mu_T) - \ln(\mu_R) < \ln(1-\varepsilon)$$

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.

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Find

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

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.

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.

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.

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.

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

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\,.$$

Alpha (Significance Level)

This option specifies one or more values for the significance level, alpha. A type-I error occurs when you reject the null hypothesis when it is true.

Values must be between zero and one. The value of 0.05 is often used for alpha. An alpha of 0.05 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.

Beta (1 - Power)

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 a false null hypothesis.

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.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Design Type	**4x2 (Balaam)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
E	**0.20**
R	**1**
N	**50 to 550 by 100**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*
COV	**0.40**

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.

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Equivalence Margin (E)	Mean Ratio for Power (R)	Coef. of Variation (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

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

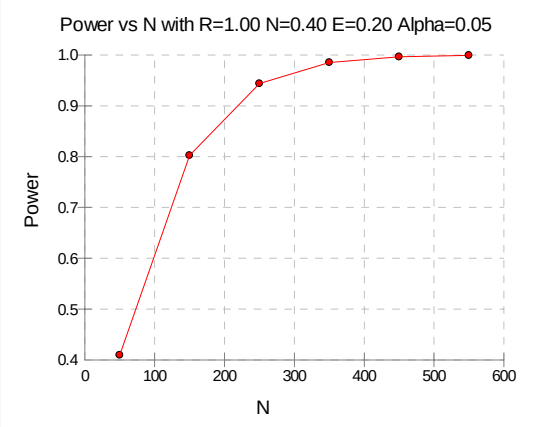
H0 (null hypothesis) is that R <= 1-E, where R = Treatment Mean / Reference Mean.
H1 (alternative hypothesis) is that R > 1-E.
Power is the probability of rejecting H0 (concluding non-inferiority) when H0 is false.
N is the total number of subjects. They are divided evenly among all sequences.
E is the 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.

Plot Section



This plot shows the power versus the sample size.

Example2 – Finding Sample Size

Continuing with Example1, the researchers want to find the exact sample size needed to achieve both 80% and 90% power.

Setup

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**4x2 (Balaam)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
E	**0.20**
R	**1**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.2 0.1**
COV	**0.40**

Annotated Output

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

Numeric Results for Testing Non-Inferiority Using the Mean Ratio
Design: Balaam's Cross-Over. Hypotheses: H0: R <= 1-E; H1: R > 1-E.

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Equivalence Margin (E)	Mean Ratio for Power (R)	Coef. of Variation (COV)	Alpha	Beta
0.9039	208	4x2	0.20	1.00	0.40	0.0500	0.0961
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.

Example3 –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 High-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

You can enter these values yourself or load the Example3 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**4x2 (Balaam)**
Higher Is	**Good**
Test Type	**Non-Inferiority**
E	**0.2**
R	**1.0**
N	*Ignored since this is the Find setting*
COV	**0.40**
Alpha	**0.05**
Beta	**0.10**

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.

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Equivalence Margin (E)	Mean Ratio for Power (R)	Coef. of Variation (COV)	Alpha	Beta
0.9039	208	4x2	0.20	1.00	0.40	0.0500	0.0961

PASS has also obtained the sample size of 208.

Chapter 540

Higher-Order Cross-Over Designs: Testing Equivalence 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

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

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

$$\mathrm{H}_0: \delta \le \varepsilon_L \text{ or } \delta \ge \varepsilon_U \text{ where } \varepsilon_L < 0,\ \varepsilon_U > 0.$$

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.

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Find

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

|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.

D (True Value)

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.

Sw (Within Std. 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.

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the significance level, alpha. A type-I error occurs when you reject the null hypothesis when it is true.

Values must be between zero and one. The value of 0.05 is often used for alpha. An alpha of 0.05 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.

Beta (1 - Power)

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 a false null hypothesis.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Design Type	**2x3 (Two-Sequence Dual)**
|EU|	**19.2**
-|EL|	**-Upper Limit**
D	**-4**
Sw	**18**
N	**4 6 8 10 12 14 16 18 20 30 40**
Alpha	**0.05**
Beta	*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 Testing the Equivalence of Two Means
Design: Two-Sequence Dual Cross-Over

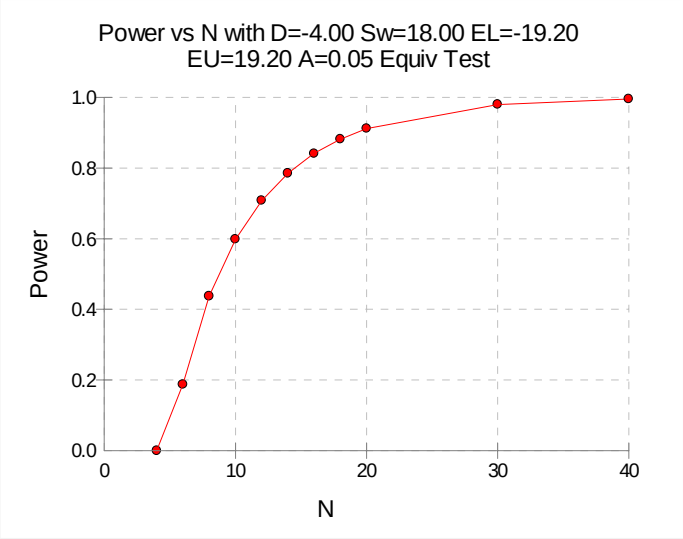
Power	Total Sample Size (N)	Sequences and Periods (SxP)	Lower Equiv. Limit (EL)	Upper Equiv. Limit (EU)	Diff. for Power (D)	Standard Error of Diff. (Sw)	Alpha	Beta
0.0000	4	2x3	-19.20	19.20	-4.00	18.00	0.0500	1.0000
0.1878	6	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.8122
0.4375	8	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.5625
0.5985	10	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.4015
0.7082	12	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.2918
0.7855	14	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.2145
0.8411	16	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.1589
0.8818	18	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.1182
0.9119	20	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.0881
0.9800	30	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.0200
0.9957	40	2x3	-19.20	19.20	-4.00	18.00	0.0500	0.0043

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. Note that 20 subjects yield about 90% power.

Plot Section



This plot shows the power versus the sample size.

Example2 – Finding Sample Size

Continuing with Example1, the researchers want to find the exact sample size needed to achieve both 80% and 90% power.

Setup

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**2x3 (Two-Sequence Dual)**
|EU|	**19.2**
-|EL|	**-Upper Limit**
D	**-4**
Sw	**18**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.10 0.20**

Annotated Output

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

Numeric Results for Testing the Equivalence of Two Means
Design: Two-Sequence Dual Cross-Over

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Lower Equiv. Limit (EL)	Upper Equiv. Limit (EU)	Diff. for Power (D)	Standard Error of Diff. (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.

Example3 – Validation using Chen

Chapter 256 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

You can enter these values yourself or load the Example3 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**4x2 (Balaam)**
|EU|	**0.2**
-|EL|	**-Upper Limit**
D	**0 0.05 0.10 0.15**
Sw	**0.1**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.10**

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: Balaam's Cross-Over

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Lower Equiv. Limit (EL)	Upper Equiv. Limit (EU)	Diff. for Power (D)	Standard Error of Diff. (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
0.9013	276	4x2	-0.20	0.20	0.15	0.10	0.0500	0.0987

PASS obtains the same samples sizes as Chen et al. (1997).

Chapter 545

Higher-Order Cross-Over Designs: Testing Equivalence 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

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

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 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

$$H_0: \phi \le \phi_L \text{ or } \phi \ge \phi_U \text{ where } \phi_L < 1, \phi_U > 1.$$

The alternative hypothesis of equivalence is

$$H_1: \phi_L < \phi < \phi_U$$

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.

$$\phi_L < \phi < \phi_U$$
$$\Rightarrow \phi_L < \left\{\frac{\mu_T}{\mu_R}\right\} < \phi_U$$
$$\Rightarrow \ln(\phi_L) < \left\{\ln(\mu_T) - \ln(\mu_R)\right\} < \ln(\phi_U)$$

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.

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

$$\mu_X = \frac{\mu_Y}{\ln\left(COV_Y^2 + 1 \right)}$$

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

$$Power = T_V\left(\left(\frac{\ln(\phi_U) - |\ln(\phi)|}{\sigma_W\sqrt{b/n}}\right) - t_{V,1-\alpha}\right) - T_V\left(t_{V,1-\alpha} - \left(\frac{-\ln(\phi_L) + |\ln(\phi)|}{\sigma_W\sqrt{b/n}}\right)\right)$$

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.

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 Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters associated with this test such as the means, sample sizes, alpha, and beta.

Find

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

RU (Upper Equiv. 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 Equiv. 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.

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.

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

$$COV_Y = \sqrt{e^{\sigma_w^2} - 1}\,.$$

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.

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*.

Alpha (Significance Level)

This option specifies one or more values for the significance level, alpha. A type-I error occurs when you reject the null hypothesis when it is true.

Values must be between zero and one. The value of 0.05 is often used for alpha. An alpha of 0.05 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.

Beta (1 - Power)

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 a false null hypothesis.

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.

Example1 – 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

You can enter these values yourself or load the Example1 template from the Template tab.

Option	Value
Data Tab	
Find	**Beta and Power**
Design Type	**2x3 (Two-Sequence Dual)**
RU	**1.25**
RL	**1/RU**
R1	**0.96**
COV	**0.40**
N	**10 20 30 40 60 80**
Alpha	**0.05**
Beta	*Ignored since this is the Find setting*

Annotated Output

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

Numeric Results

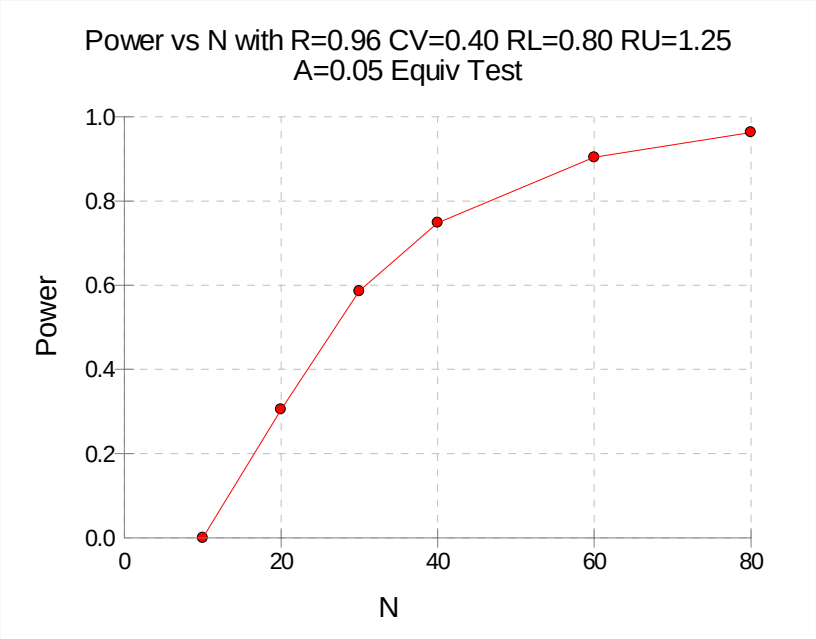
Power	Total Sample Size (N)	Sequences and Periods (SxP)	Lower Equiv. Limit (RL)	Upper Equiv. Limit (RU)	Mean Ratio for Power (R1)	Coef. of Variation (COV)	Alpha	Beta
0.0000	10	2x3	0.80	1.25	0.96	0.40	0.0500	1.0000
0.3051	20	2x3	0.80	1.25	0.96	0.40	0.0500	0.6949
0.5858	30	2x3	0.80	1.25	0.96	0.40	0.0500	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%.

Plot Section



This plot shows the power versus the sample size.

Example2 – 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

You can enter these values yourself or load the Example2 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal per Sequence)**
Design Type	**2x3 (Two-Sequence Dual)**
RU	**1.25**
RL	**1/RU**
R1	**0.96**
COV	**0.40**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.10 0.20**

Annotated Output

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

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Lower Equiv. Limit (RL)	Upper Equiv. Limit (RU)	Mean Ratio for Power (R1)	Coef. of Variation (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.

Example3 – 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

You can enter these values yourself or load the Example3 template from the Template tab.

Option	Value
Data Tab	
Find	**N (Equal Per Sequence)**
Design Type	**4x2 (Balaam)**
RU	**1.25**
RL	**1/RU**
R	**1**
COV	**0.10025**
N	*Ignored since this is the Find setting*
Alpha	**0.05**
Beta	**0.10 0.20**

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 Using Ratios
Design: Balaam's Cross-Over

Power	Total Sample Size (N)	Sequences and Periods (SxP)	Lower Equiv. Limit (RL)	Upper Equiv. Limit (RU)	Mean Ratio for Power (R)	Coef. of Variation (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).

References

Agresti, A. and Coull, B. 1998. “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” *American Statistician*, Volume 52 Number 2, pages 119-126.

A’Hern, R. P. A. 2001. "Sample size tables for exact single-stage phase II designs." *Statistics in Medicine*, Volume 20, pages 859-866.

AIAG (Automotive Industry Action Group). 1995. *Measurement Systems Analysis*. This booklet was developed by Chrysler/Ford/GM Supplier Quality Requirements Task Force. It gives a detailed discussion of how to design and analyze an R&R study. The book may be obtained from ASQC or directly from AIAG by calling 801-358-3570.

Albert, A. and Harris, E. 1987. *Multivariate Interpretation of Clinical Laboratory Data*. Marcel Dekker, New York, New York. This book is devoted to a discussion of how to apply multinomial logistic regression to medical diagnosis. It contains the algorithm that is the basis of our multinomial logistic regression routine.

Allen, D. and Cady, F.. 1982. *Analyzing Experimental Data by Regression*. Wadsworth. Belmont, Calif. This book works completely through several examples. It is very useful to those who want to see complete analyses of complex data.

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.

Altman, Douglas. 1991. *Practical Statistics for Medical Research*. Chapman & Hall. New York, NY. This book provides an introductory discussion of many statistical techniques that are used in medical research. It is the only book we found that discussed ROC curves.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. 1997. *Statistical Models Based on Counting Processess*. Springer-Verlag, New York. This is an advanced book giving many of the theoretically developments of survival analysis.

Anderson, R.L. and Hauck, W.W. 1983. "A new Procedure for testing equivalence in comparative bioavailability and other clinical trials." *Commun. Stat. Theory Methods.*, Volume 12, pages 2663-2692.

Anderson, T.W. and Darling, D.A. 1954. "A test of goodness-of-fit." *J. Amer. Statist. Assoc*, Volume 49, pages 765-769.

Andrews, D.F., and Herzberg, A.M. 1985. *Data*. Springer-Verlag, New York. This book is a collection of many different data sets. It gives a complete description of each.

Armitage. 1955. “Tests for linear trends in proportions and frequencies.” *Biometrics*, Volume 11, pages 375-386.

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.

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.

Baker, Frank. 1992. *Item Response Theory*. Marcel Dekker. New York. This book contains a current overview of IRT. It goes through the details, providing both formulas and computer code. It is not light reading, but it will provide you with much of what you need if you are attempting to use this technique.

Barnard, G.A. 1947. "Significance tests for 2 x 2 tables." *Biometrika* 34:123-138.

Barrentine, Larry B. 1991. *Concepts for R&R Studies*. ASQC Press. Milwaukee, Wisconsin. This is a very good applied work book on the subject of repeatability and reproducibility studies. The ISBN is 0-87389-108-2. ASQC Press may be contacted at 800-248-1946.

Bartholomew, D.J. 1963. "The Sampling Distribution of an Estimate Arising in Life Testing." *Technometrics*, Volume 5 No. 3, 361-374.

Bartlett, M.S. 1950. "Tests of significance in factor analysis." *British Journal of Psychology (Statistical Section)*, 3, 77-85.

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.

Belsley, Kuh, and Welsch. 1980. *Regression Diagnostics*. John Wiley & Sons. New York. This is the book that brought regression diagnostics into the main-stream of statistics. It is a graduate level treatise on the subject.

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.

Bloomfield, P. 1976. *Fourier Analysis of Time Series*. John Wiley and Sons. New York. This provides a technical introduction to fourier analysis techniques.

Bock, R.D., Aiken, M. 1981. "Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, 46, 443-459.

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. and Jenkins, G.M. 1976. *Time Series Analysis - Forecasting and Control*. Holden-Day.: San Francisco, California. This is the landmark book on ARIMA time series analysis. Most of the material in chapters 6 - 9 of this manual comes from this work.

Box, G.E.P. 1949. "A general distribution theory for a class of likelihood criteria." *Biometrika*, 1949, **36**, 317-346.

Box, G.E.P. 1954a. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: I." *Annals of Mathematical Statistics*, **25**, 290-302.

Box, G.E.P. 1954b. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: II." *Annals of Mathematical Statistics*, **25**, 484-498.

Box, G.E.P., Hunter, S. and Hunter. 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.

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.

Burdick, R.K. and Larsen, G.A. 1997. "Confidence Intervals on Measures of Variability in R&R Studies." *Journal of Quality Technology, Vol. 29, No. 3, Pages 261-273*. This article presents the formulas used to construct confidence intervals in an R&R study.

Bury, Karl. 1999. *Statistical Distributions in Engineering.*. Cambridge University Press. New York, NY. (www.cup.org).

Cameron, A.C. and Trivedi, P.K. 1998. *Regression Analysis of Count Data*. Cambridge University Press. New York, NY. (www.cup.org).

Carmines, E.G. and Zeller, R.A. 1990. *Reliability and Validity Assessment.* Sage University Paper. 07-017. Newbury Park, CA.

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.

Cattell, R.B. 1966. "The scree test for the number of factors." *Mult. Behav. Res.* 1, 245-276.

Cattell, R.B. and Jaspers, J. 1967. "A general plasmode (No. 30-10-5-2) for factor analytic exercises and research." *Mult. Behav. Res. Monographs*. 67-3, 1-212.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. 1983. *Graphicals Methods for Data Analysis.* Duxbury Press, Boston, Mass. This wonderful little book is full of examples of ways to analyze data graphically. It gives complete (and readable) coverage to such topics as scatter plots, probability plots, and box plots. It is strongly recommended.

Chatfield, C. 1984. *The Analysis of Time Series*. Chapman and Hall. New York. This book gives a very readable account of both ARMA modeling and spectral analysis. We recommend it to those who wish to get to the bottom of these methods.

Chatterjee and Price. 1979. *Regression Analysis by Example*. John Wiley & Sons. New York. A great hands-on book for those who learn best from examples. A newer edition is now available.

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.

Chen, Xun. 2002. "A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases." *Statistics in Medicine*, Volume 21, pages 943-956.

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.

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.

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.

Conover, W.J. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons, Inc. New York.

Conover, W.J., Johnson, M.E., and **Johnson, M.M.** 1981. *Technometrics*, **23,** 351-361.

Cook, D. and Weisberg, S. 1982. *Residuals and Influence in Regression*. Chapman and Hall. New York. This is an advanced text in the subject of regression diagnostics.

Cooley, W.W. and Lohnes, P.R. 1985. *Multivariate Data Analysis.* Robert F. Krieger Publishing Co. Malabar, Florida.

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.

Cox, D. R. 1975. "Contribution to discussion of Mardia (1975a)." *Journal of the Royal Statistical Society, Series B*, Volume 37, Pages 380-381.

Cureton, E.E. and D'Agostino, R.B. 1983. *Factor Analysis - An Applied Approach*. Lawrence Erlbaum Associates. Hillsdale, New Jersey. (This is a wonderful book for those who want to learn the details of what factor analysis does. It has both the theoretical formulas and simple worked examples to make following along very easy.)

D'Agostino, R.B., Belanger, A., D'Agostino, R.B. Jr. 1990."A Suggestion for Using Powerful and Informative Tests of Normality.", *The American Statistician*, November 1990, Volume 44 Number 4, pages 316-321. This tutorial style article discusses D'Agostino's tests and tells how to interpret normal probability plots.

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.

Dallal, G. 1986. "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality," *The American Statistician*, Volume 40, Number 4, pages 294-296.

Daniel, C. and Wood, F. 1980. *Fitting Equations to Data*. John Wiley & Sons. New York. This book gives several in depth examples of analyzing regression problems by computer.

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.

Davis, J. C. 1985. *Statistics and Data Analysis in Geology*. John Wiley. New York. (A great layman's discussion of many statistical procedures, including factor analysis.)

Davison, A.C. and Hinkley, D.V. 1999. *Bootstrap Methods and their Applications*. Cambridge University Press. NY, NY. This book provides and detailed account of bootstrapping.

Davison, Mark. 1983. *Multidimensional Scaling*. John Wiley & Sons. NY, NY. This book provides a very good, although somewhat advanced, introduction to the subject.

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.

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. Academic Press. New York. (Presents many useful results for determining sample sizes.)

DeVor, Chang, and Sutherland. 1992. *Statistical Quality Design and Control*. Macmillan Publishing. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 800 pages.

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.

Dillon, W. and Goldstein, M. 1984. *Multivariate Analysis - Methods and Applications*. John Wiley. NY, NY. This book devotes a complete chapter to loglinear models. It follows Fienberg's book, providing additional discussion and examples.

Dixon, W. J. and Tukey, J. W. 1968. "Approximate behavior of the distribution of Winsorized t," *Technometrics*, Volume 10, pages 83-98.

Dodson, B. 1994. *Weibull Analysis*. ASQC Quality Press. Milwaukee, Wisconsin. This paperback book provides the basics of Weibull fitting. It contains many of the formulas used in our Weibull procedure.

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.

Draper, N.R. and Smith, H. 1966. *Applied Regression Analysis*. John Wiley & Sons. New York. This is a classic text in regression analysis. It contains both in depth theory and applications. This text is often used in graduate courses in regression analysis.

Draper, N.R. and Smith, H. 1981. *Applied Regression Analysis - Second Edition*. John Wiley & Sons. New York, NY. This is a classic text in regression analysis. It contains both in-depth theory and applications. It is often used in graduate courses in regression analysis.

du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H. 1986. *Graphical Exploratory Data Analysis.* Springer-Verlag. New York. This book contains examples of graphical analysis for a broad range of topics.

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.

Dunteman, G.H. 1989. *Principal Components Analysis*. Sage University Papers, 07-069. Newbury Park, California. Telephone (805) 499-0721. This monograph costs only $7. It gives a very good introduction to PCA.

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.

Durbin, J. and Watson, G. S. 1950. "Testing for Serial Correlation in Least Squares Regression - I," *Biometrika*, Volume 37, pages 409-428.

Durbin, J. and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression - II," *Biometrika*, Volume 38, pages 159-177.

Dyke, G.V. and Patterson, H.D. 1952. "Analysis of factorial arrangements when the data are proportions." *Biometrics*. Volume 8, pages 1-12. This is the source of the data used in the LLM tutorial.

Eckert, Joseph K. 1990. *Property Appraisal and Assessment Administration*. International Association of Assessing Officers. 1313 East 60th Street. Chicago, IL 60637-2892. Phone: (312) 947-2044. This is a how-to manual published by the IAAO that describes how to apply many statistical procedures to real estate appraisal and tax assessment. We strongly recommend it to those using our *Assessment Model* procedure.

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. 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.

Everitt, B.S. and Dunn, G. 1992. *Applied Multivariate Data Analysis*. Oxford University Press. New York. This book provides a very good introduction to several multivariate techniques. It helps you understand how to interpret the results.

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.; & Salih, F.A. 1987. "Statistical inference for coefficient alpha." *Applied Psychological Measurement*, Vol. 11, pages 93-103.

Feldt, L.S.; 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.

Fienberg, S. 1985. *The Analysis of Cross-Classified Categorical Data*. MIT Press. Cambridge, Massachusetts. This book provides a very good introduction to the subject. It is a must for any serious student of the subject.

Finney, D. 1971. *Probit Analysis*. Cambridge University Press. New York, N.Y.

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge University Press. New York, New York.

Fisher, R.A. 1936. "The use of multiple measurements in taxonomic problems." *Annuals of Eugenics*, Volume 7, Part II, 179-188. This article is famous because in it Fisher included the 'iris data' that is always presented when discussing discriminant analysis.

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, 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.

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.

Fleming, T. R. 1982. "One-sample multiple testing procedure for Phase II clinical trials." *Biometrics*, Volume 38, pages 143-151.

Flury, B. and Riedwyl, H. 1988. *Multivariate Statistics: A Practical Approach*. Chapman and Hall. New York. This is a short, paperback text that provides lots of examples.

Flury, B. 1988. *Common Principal Components and Related Multivariate Models*. John Wiley & Sons. New York. This reference describes several advanced PCA procedures.

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.

Gleason, T.C. and Staelin, R. 1975. "A proposal for handling missing data." *Psychometrika*, 40, 229-252.

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.

Gomez, K.A. and Gomez, A. A. 1984. *Statistical Procedures for Agricultural Research*. John Wiley & Sons. New York. This reference contains worked-out examples of many complex ANOVA designs. It includes split-plot designs. We recommend it.

Graybill, Franklin. 1961. *An Introduction to Linear Statistical Models*. McGraw-Hill. New York, New York. This is an older book on the theory of linear models. It contains a few worked examples of power analysis.

Greenacre, M. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press. Orlando, Florida. This book goes through several examples. It is probably the most complete book in English on the subject.

Greenacre, Michael J. 1993. *Correspondence Analysis in Practice*. Academic Press. San Diego, CA. This book provides a self-teaching course in correspondence analysis. It is the clearest exposition on the subject that I have every seen. If you want to gain an understanding of CA, you must obtain this (paperback) book.

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.

Haberman, S.J. 1972. "Loglinear Fit of Contingency Tables." *Applied Statistics*. Volume 21, pages 218-225. This lists the fortran program that is used to create our LLM algorithm.

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

Hambleton, R.K; Swaminathan, H; Rogers, H.J. 1991. *Fundamentals of Item Response Theory*. Sage Publications. Newbury Park, California. Phone: (805)499-0721. Provides an inexpensive, readable introduction to IRT. A good place to start.

Hamilton, L. 1991. *Regression with Graphics: A Second Course in Applied Statistics*. Brooks/Cole Publishing Company. Pacific Grove, California. This book gives a great introduction to the use of graphical analysis with regression. It is a must for any serious user of regression. It is written at an introductory level.

Hand, D.J. and Taylor, C.C. 1987. *Multivariate Analysis of Variance and Repeated Measures*. Chapman and Hall. London, England.

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.

Hartigan, J. 1975. *Clustering Algorithms*. John Wiley. New York. (This is the "bible" of cluster algorithms. Hartigan developed the K-means algorithm used in **NCSS**.)

Haupt, R.L. and Haupt, S.E. 1998. *Practical Genetic Algorithms*. John Wiley. New York.

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.

Hintze, J. L. and Nelson, R.D. 1998. "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician* 52, 181-184.

Hoaglin, Mosteller, and Tukey. 1985. *Exploring Data Tables, Trends, and Shapes*. John Wiley. New York.

Hoaglin, Mosteller, and Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons. New York.

Hoerl, A.E. and Kennard, R.W. 1970. "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 55-82.

Hoerl, A.E. and Kennard R.W. 1976. "Ridge regression: Iterative estimation of the biasing parameter." *Communications in Statistics* A5, 77-88.

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.

Jackson, J.E. 1991. *A User's Guide To Principal Components.* John Wiley & Sons. New York. This is a great book to learn about PCA from. It provides several examples and treats everything at a level that is easy to understand.

James, Mike. 1985. *Classification Algorithms*. John Wiley & Sons. New York. This is a great text on the application of discriminant analysis. It includes a simple, easy-to-understand, theoretical development as well as discussions of the application of discriminant analysis.

Jammalamadaka, S.R. and SenGupta, A. 2001. *Topics in Circular Statistics*. World Scientific. River Edge, New Jersey.

Jobson, J.D. 1992. *Applied Multivariate Data Analysis - Volume II: Categorical and Multivariate Methods*. Springer-Verlag. New York. This book is a useful reference for loglinear models and other multivariate methods. It is easy to follows and provides lots of examples.

Jolliffe, I.T. 1972. "Discarding variables in a principal component analysis, I: Artifical data." *Applied Statistics*, 21:160-173.

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.

Jolliffe, I.T. 1986. *Principal Component Analysis*. Springer-Verlag. New York. This book provides an easy-reading introduction to PCA. It goes through several examples.

Julious, Steven A. 2004. "Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data." *Statistics in Medicine*, 23:1921-1986.

Juran, J.M. 1979. *Quality Control Handbook*. McGraw-Hill. New York.

Kaiser, H.F. 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement*. 20:141-151.

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.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data*. John Wiley. New York. This book gives an excellent introduction to cluster analysis. It treats the forming of the distance matrix and several different types of cluster methods, including fuzzy. All this is done at an elementary level so that users at all levels can gain from it.

Kay, S.M. 1988. *Modern Spectral Estimation*. Prentice-Hall: Englewood Cliffs, New Jersey. A very technical book on spectral theory.

Kendall,M. and Ord, J.K. 1990. *Time Series.* Oxford University Press. New York. This is theoretical introduction to time series analysis that is very readable.

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.

Keppel, Geoffrey. 1991. *Design and Analysis - A Researcher's Handbook.* Prentice Hall. Englewood Cliffs, New Jersey. This is a very readable primer on the topic of analysis of variance. Recommended for those who want the straight scoop with a few, well-chosen examples.

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.

Kotz and Johnson. 1993. *Process Capability Indices.* Chapman & Hall. New York. This book gives a detailed account of the capability indices used in SPC work. 207 pages.

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.

Kruskal, J. 1964. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika* 29, pages 1-27, 115-129. This article presents the algorithm on which the non-metric algorithm used in NCSS is based.

Kruskal, J. and Wish, M. 1978. *Multidimensional Scaling*. Sage Publications. Beverly Hills, CA. This is a well-written monograph by two of the early pioneers of MDS. We suggest it to all serious students of MDS.

Lachenbruch, P.A. 1975. *Discriminant Analysis.* Hafner Press. New York. This is an in-depth treatment of the subject. It covers a lot of territory, but has few examples.

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.

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.

Lebart, Morineau, and Warwick. 1984. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons. This book devotes a large percentage of its discussion to correspondence analysis.

Lee, E.T. 1974. "A Computer Program for Linear Logistic Regression Analysis" in *Computer Programs in Biomedicine,* Volume 4, pages 80-92.

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, S. K. 1977. "On the Asymptotic Variances of u Terms in Loglinear Models of Multidimensional Contingency Tables." *Journal of the American Statistical Association*. Volume 72 (June, 1977), page 412. This article describes methods for computing standard errors that are used in the LLM section of this program.

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.

Lesaffre, E. and Albert, A. 1989. "Multiple-group Logistic Regression Diagnostics" *Applied Statistics*, Volume 38, pages 425-440. See also Pregibon 1981.

Levene, H. 1960. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al., eds. Stanford University Press, Stanford Calif., pp. 278-292.

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.

Little, R. and Rubin, D. 1987. *Statistical Analysis with Missing Data*. John Wiley & Sons. New York. This book is completely devoted to dealing with missing values. It gives a complete treatment of using the EM algorithm to estimate the covariance matrix.

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.

Lui, 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.

Lloyd, D.K. and Lipow, M. 1991. *Reliability: Management, Methods, and Mathematics*. ASQC Quality Press. Milwaukee, Wisconsin.

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.

Lockhart, R. A. & Stephens, M. A. 1985. "Tests of fit for the von Mises distribution." *Biometrika* 72, pages 647-652.

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.

Makridakis, S. and Wheelwright, S.C. 1978. *Iterative Forecasting*. Holden-Day.: San Francisco, California. This is a very good book for the layman since it includes several detailed examples. It is written for a person with a minimum amount of mathematical background.

Manly, B.F.J. 1986. *Multivariate Statistical Methods - A Primer*. Chapman and Hall. New York. This nice little paperback provides a simplified introduction to many multivariate techniques, including MDS.

Mardia, K.V. and Jupp, P.E. 2000. *Directional Statistics*. John Wiley & Sons. New York.

Marple, S.L. 1987. *Digital Spectral Analysis with Applications*. Prentice-Hall: Englewood Cliffs, New Jersey. A technical book about spectral analysis.

Martinez and Iglewicz. 1981. "A test for departure from normality based on a biweight estimator of scale." *Biometrika*, 68, 331-333).

Marubini, E. and **Valsecchi, M.G.** 1996. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley: New York, New York.

Mather, Paul. 1976. *Computational Methods of Multivariate Analysis in Physical Geography.* John Wiley & Sons. This is a great book for getting the details on several multivariate procedures. It was written for non-statisticians. It is especially useful in its presentation of cluster analysis. Unfortunately, it is out-of-print. You will have to look for it in a university library (it is worth the hunt).

Matsumoto, M. and Nishimura,T. 1998. "Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator" *ACM Trans. On Modeling and Computer Simulations*.

Mauchly, J.W. 1940. "Significance test for sphericity of a normal n-variate distribution." *Annals of Mathematical Statistics*, 11: 204-209

McCabe, G.P. 1984. "Principal variables." *Technometrics*, 26, 137-144.

McClish, D.K. 1989. "Analyzing a Portion of the ROC Curve." *Medical Decision Making*, 9: 190-195

McHenry, Claude. 1978. "Multivariate subset selection." *Journal of the Royal Statistical Society, Series C*. Volume 27, No. 23, pages 291-296.

McNeil, D.R. 1977. *Interactive Data Analysis*. John Wiley & Sons. New York.

Mendenhall, W. 1968. *Introduction to Linear Models and the Design and Analysis of Experiments.* Wadsworth. Belmont, Calif.

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.

Milliken, G.A. and Johnson, D.E. 1984. *Analysis of Messy Data, Volume 1*. Van Nostrand Rienhold. New York, NY.

Milne, P. 1987. *Computer Graphics for Surveying.* E. & F. N. Spon, 29 West 35th St., NY, NY 10001

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.

Montgomery, Douglas and Peck. 1992. *Introduction to Linear Regression Analysis.* A very good book on this topic.

Montgomery, Douglas C. 1991. *Introduction to Statistical Quality Control.* Second edition. John Wiley & Sons. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 700 pages.

Moore, D. S. and McCabe, G. P. 1999. *Introduction to the Practice of Statistics*. W. H. Freeman and Company. New York.

Mosteller, F. and Tukey, J.W. 1977. *Data Analysis and Regression*. Addison-Wesley. Menlo Park, California. This book should be read by all serious users of regression analysis. Although the terminology is a little different, this book will give you a fresh look at the whole subject.

Motulsky, Harvey. 1995. *Intuitive Biostatistics.* Oxford University Press. New York, New York. This is a wonderful book for those who want to understand the basic concepts of statistical testing. The author presents a very readable coverage of the most popular biostatistics tests. If you have forgotten how to interpret the various statistical tests, get this book!

Moura, Eduardo C. 1991. *How To Determine Sample Size And Estimate Failure Rate in Life Testing*. ASQC Quality Press. Milwaukee, Wisconsin.

Mueller, 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.

Mueller, 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.

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.

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.

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.

Nash, J. C. 1987. *Nonlinear Parameter Estimation*. Marcel Dekker, Inc. New York, NY.

Nash, J.C. 1979. *Compact Numerical Methods for Computers*. John Wiley & Sons. New York, NY.

Nel, D.G. and van der Merwe, C.A. 1986. "A solution to the multivariate Behrens-Fisher problem." *Communications in Statistics—Series A, Theory and Methods,* 15, pages 3719-3735.

Nelson, W.B. 1982. *Applied Life Data Analysis*. John Wiley, New York.

Nelson, W.B. 1990. *Accelerated Testing*. John Wiley, New York.

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.

Newton, H.J. 1988. *TIMESLAB: A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole: Pacific Grove, California. This book is loaded with theoretical information about time series analysis. It includes software designed by Dr. Newton for performing advanced time series and spectral analysis. The book requires a strong math and statistical background.

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).

Orloci, L. & Kenkel, N. 1985. *Introduction to Data Analysis*. International Co-operative Publishing House. Fairland, Maryland. This book was written for ecologists. It contains samples and BASIC programs of many statistical procedures. It has one brief chapter on MDS, and it includes a non-metric MDS algorithm.

Ostle, B. 1988. *Statistics in Research. Fourth Edition*. Iowa State Press. Ames, Iowa. A comprehension book on statistical methods.

Ott, L. 1977. *An Introduction to Statistical Methods and Data Analysis*. Wadsworth. Belmont, Calif. Use the second edition.

Ott, L. 1984. *An Introduction to Statistical Methods and Data Analysis, Second Edition*. Wadsworth. Belmont, Calif. This is a complete methods text. Regression analysis is the focus of five or six chapters. It stresses the interpretation of the statistics rather than the calculation, hence it provides a good companion to a statistical program like ours.

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.

Pandit, S.M. and Wu, S.M. 1983. *Time Series and System Analysis with Applications*. John Wiley and Sons. New York. This book provides an alternative to the Box-Jenkins approach for dealing with ARMA models. We used this approach in developing our automatic ARMA module.

Parmar, M.K.B. and Machin, D. 1995. *Survival Analysis*. John Wiley and Sons. New York.

Parmar, M.K.B., Torri, V., and Steart, L. 1998. "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints." *Statistics in Medicine* 17, 2815-2834.

Pearson, K. 1901. "On lines and planes of closest fit to a system of points in space." *Philosophical Magazine* 2, 557-572.

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.

Pedhazur, E.L. and Schmelkin, L.P. 1991. *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. This mammoth book (over 800 pages) covers multivariate analysis, regression analysis, experimental design, analysis of variance, and much more. It provides annotated output from SPSS and SAS which is also useful to our users. The text is emphasizes the social sciences. It provides a "how-to," rather than a theoretical, discussion. Its chapters on factor analysis are especially informative.

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.

Pocock, S.J. 1977. "Group sequential methods in the design and analysis of clinical trials." *Biometrika*, 64, pages 191-199.

Press, S. J. and Wilson, S. 1978. "Choosing Between Logistic Regression and Discriminant Analysis." *Journal of the American Statistical Association*, Volume 73, Number 364, Pages 699-705. This article details the reasons why logistic regression should be the preferred technique.

Press, William H. 1986. *Numerical Recipes*, Cambridge University Press, New York, New York.

Pregibon, Daryl. 1981. "Logistic Regression Diagnostics." *Annals of Statistics*, Volume 9, Pages 705-725. This article details the extensions of the usual regression diagnostics to the case of logistic regression. These results were extended to multiple-group logistic regression in Lesaffre and Albert (1989).

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.

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.

Ratkowsky, David A. 1989. *Handbook of Nonlinear Regression Models*. Marcel Dekker. New York. A good, but technical, discussion of various nonlinear regression models.

Rawlings John O. 1988. *Applied Regression Analysis: A Research Tool*. Wadsworth. Belmont, California. This is a readable book on regression analysis. It provides a thorough discourse on the subject.

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.

Ryan, Thomas P. 1989. *Statistical Methods for Quality Improvement*. John Wiley & Sons. New York. This is a comprehensive treatment of SPC including control charts, process capability, and experimental design. It provides many rules-of-thumb and discusses many non-standard situations. This is a very good 'operators manual' type of book. 446 pages.

Ryan, Thomas P. 1997. *Modern Regression Methods*. John Wiley & Sons. New York. This is a comprehensive treatment of regression analysis. The author often deals with practical issues that are left out of other texts.

Sahai, Hardeo & Khurshid, Anwer. 1995. *Statistics in Epidemiology*. CRC Press. Boca Raton, Florida.

Schiffman, Reynolds, & Young. 1981. *Introduction to Multidimensional Scaling*. Academic Press. Orlando, Florida. This book goes through several examples.

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.

Schmee and Hahn. November, 1979. "A Simple Method for Regression Analysis." *Technometrics*, Volume 21, Number 4, pages 417-432.

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.

Seber, G.A.F. 1984. *Multivariate Observations*. John Wiley & Sons. New York. (This book is an encyclopedia of multivariate techniques. It emphasizes the mathematical details of each technique and provides a complete set of references. It will only be useful to those comfortable with reading mathematical equations based on matrices.)

Seber, G.A.F. and Wild, C.J. 1989. *Nonlinear Regression*. John Wiley & Sons. New York. This book is an encyclopedia of nonlinear regression.

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.

Shapiro, S.S. and Wilk, M.B. 1965 "An analysis of Variance test for normality." *Biometrika*, Volume 52, pages 591-611.

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.

Snedecor, G. and Cochran, Wm. 1972. *Statistical Methods*. The Iowa State University Press. Ames, Iowa.

Sorribas, A., March, J., and Trujillano, J. 2002. "A new parametric method based on S-distributions for computing receiver operating characteristic curves for continuous diagnostic tests." *Statistics in Medicine* 21, 1213-1235.

Spath, H. 1985. *Cluster Dissection and Analysis.* Halsted Press. New York. (This book contains a detailed discussion of clustering techniques for large data sets. It contains some heavy mathematical notation.)

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Song, F. 2000. *Methods for Meta-Analysis in Medical Research.* John Wiley & Sons. New York.

Swets, John A. 1996. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics - Collected Papers.* Lawrence Erlbaum Associates. Mahway, New Jersey.

Tabachnick, B. and Fidell, L. 1989. *Using Multivariate Statistics.* Harper Collins. 10 East 53d Street, NY, NY 10022. This is an extremely useful text on multivariate techniques. It presents computer printouts and discussion from several popular programs. It provides checklists for each procedure as well as sample written reports. I strongly encourage you to obtain this book!

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.

Thomopoulos, N.T. 1980. *Applied Forecasting Methods*. Prentice-Hall: Englewood Cliffs, New Jersey. This book contains a very good presentation of the classical forecasting methods discussed in chapter two.

Thompson, Simon G. 1998. *Encyclopedia of Biostatistics, Volume 4*. John Wiley & Sons. New York. Article on Meta-Analysis on pages 2570-2579.

Tiku, M. L. 1965. "Laguerre Series Forms of Non-Central X^2 and F Distributions," *Biometrika*, Volume 42, pages 415-427.

Torgenson, W.S. 1952. “Multidimensional scaling. I. Theory and method.” *Psychometrika* 17, 401-419. This is one of the first articles on MDS. There have been many advances, but this article presents many insights into the application of the technique. It describes the algorithm on which the metric solution used in this program is based.

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.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Company. Reading, Mass.

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.

Upton, G.J.G. and Fingleton, B. 1989. *Spatial Data Analysis by Example: Categorical and Directional Data. Volume 2.* John Wiley & Sons. New York.

Velicer, W.F. 1976. "Determining the number of components from the matrix of partial correlations." *Psychometrika*, 41, 321-327.

Velleman, Hoaglin. 1981. *ABC's of Exploratory Data Analysis*. Duxbury Press, Boston, Massachusets.

Voit, E.O. 1992. "The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions." *Biometrical J.* 34, 855-878.

Voit, E.O. 2000. "A Maximum Likelihood Estimator for Shape Parameters of S-Distributions." *Biometrical J.* 42, 471-479.

Voit, E.O. and Schwacke, L. 1998. "Scalability properties of the S-distribution." *Biometrical J.* 40, 665-684.

Voit, E.O. and Yu, S. 1994. "The S-distribution. Approximation of discrete distributions." *Biometrical J.* 36, 205-219.

Walter, S.D., Eliasziw, M., and Donner, A. 1998. "Sample Size and Optimal Designs For Reliability Studies.” *Statistics in Medicine*, 17, 101-110.

Welch, B.L. 1938. "The significance of the difference between two means when the population variances are unequal.” *Biometrika*, 29, 350-362.

Westgard, J.O. 1981. “A Multi-Rule Shewhart Chart for Quality Control in Clinical Chemistry,” *Clinical Chemistry*, Volume 27, No. 3, pages 493-501. (This paper is available online at the www.westgard.com).

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.

Wickens, T.D. 1989. *Multiway Contingency Tables Analysis for the Social Sciences.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. A thorough book on the subject. Discusses loglinear models in depth.

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.

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.

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.

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

Note that most index entries are of the form “chapter-page”. If no chapter is given, the entry refers to the introductory chapters. A list of chapters is given in the Table of Contents.