

# **User's Guide - I**

**Quick Start, Proportions, and ROC Curves**

**PASS**  
**Power Analysis and Sample Size System**

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

# PASS User's Guide - I

Copyright © 2006  
Dr. Jerry L. Hintze  
Kaysville, Utah 84037

All Rights Reserved  
Printed in the United States of America

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.

## Warning:

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

## About this manual

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

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

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

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

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

**Jerry Hintze, Author**

# PASS License Agreement

*Important: The enclosed Power Analysis and Sample Size software program (PASS) is licensed by NCSS to customers for their use only on the terms set forth below. 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 (Power Analysis and Sample Size)** 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](http://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

## User's Guide - I

### Quick Start

- 1 Installation 1
- 2 Running PASS 7
- 3 PASS Home 11
- 4 Procedure Window 15
- 5 Output Window 33
- 6 Introduction to Power Analysis 41
- 7 Introduction to Proportions 51
- 8 Introduction to Means 55

### Proportions

#### One Proportion

- 100 Inequality
- 105 Non-Inferiority
- 110 Equivalence
- 115 Confidence Interval
- 120 Single-Stage Design
- 125 Two-Stage Design
- 130 Three-Stage Design
- 135 Post-Marketing Surveillance

#### Two Correlated Proportions

- 150 Inequality
- 155 Matched Case-Control
- 160 Non-Inferiority
- 165 Equivalence

#### Two Independent Proportions

- 200 Inequality
- 205 Inequality (Offset)
- 210 Non-Inferiority
- 215 Equivalence
- 220 Group Sequential
- 225 Mantel-Haenszel

#### Cluster-Randomized

- 230 Inequality
- 235 Non-Inferiority
- 240 Equivalence

#### Many Proportions

- 250 Chi-Square Test

### ROC Curves

- 260 One ROC Curve
- 265 Two ROC Curves

### References

#### Index

## User's Guide - II

### Means

#### One-Mean

- 400 Inequality (Normal)
- 405 Inequality (Exponential)
- 410 Inequality (Simulation)
- 415 Non-Inferiority
- 420 Confidence Interval

#### Two Independent Means

- 430 Inequality (Normal)
- 435 Inequality (Exponential)
- 440 Inequality (Simulation)
- 445 Inequality – Ratios
- 450 Non-Inferiority – Differences
- 455 Non-Inferiority – Ratios
- 460 Equivalence
- 465 Equivalence (Simulation)
- 470 Equivalence – Ratios
- 475 Group Sequential
- 480 Cluster Randomization

#### Two Correlated Means

- 490 Inequality (Simulation)
- 495 Equivalence (Simulation)

#### 2x2 Cross-Over

- 500 Inequality – Differences
- 505 Inequality – Ratios
- 510 Non-Inferiority – Differences
- 515 Non-Inferiority – Ratios
- 520 Equivalence – Differences
- 525 Equivalence – Ratios

#### Higher-Order Cross-Over

- 530 Non-Inferiority – Differences
- 535 Non-Inferiority – Ratios
- 540 Equivalence – Differences
- 545 Equivalence – Ratios

### References

#### Index

## User's Guide - III

### Many Means - ANOVA

- 550 One-Way
- 555 One-Way (Simulation)
- 560 Factorial
- 565 Randomized Block
- 570 Repeated Measures

### Multiple Comparisons

- 575 Analytic
- 580 Pair-Wise (Simulation)
- 585 Treatment / Control (Simulation)
- 590 Contrast (Simulation)

### Multivariate Routines

- 600 Hotelling's T-Squared
- 605 MANOVA

### Simulation

- 630 Data Simulator

### Variances

- 650 One Variance
- 655 Two Variances

### Survival Analysis

- 700 Simple Log Rank
- 705 Advanced Log Rank
- 710 Group Sequential Log Rank

### Correlations

- 800 One Correlation
- 805 Two Correlations
- 810 Intraclass Correlation
- 815 Coefficient Alpha: 1
- 820 Coefficient Alpha: 2

### Regression

- 850 Cox Regression
- 855 Linear Regression
- 860 Logistic Regression
- 865 Multiple Regression
- 870 Poisson Regression

### Helps and Aids

- 900 Chi-Square Effect Size
- 905 Standard Deviation
- 910 Odds Ratio

### References

#### Index



## CHAPTER 1

# Installation

## Before you install

### 1. Check system requirements

*PASS* runs on 32-bit Windows systems. This includes Windows 98, Windows ME, Windows NT 4.0, Windows 2000, and Windows XP. The recommended minimum system is a Pentium PC with 64 MB of memory.

*PASS* takes up about 80 MB of disk space. Once installed, *PASS* also requires about 20 MB of temporary disk space while it is running.

### 2. Find a home for *PASS*

Before you start installing, decide on a folder where you want to install *PASS*. By default, the setup program will install *PASS* in the NCSS97 (or NCSS2000) folder of your C drive. You may change this during the installation, but not after.

### 3. If you have a previous NCSS

*PASS* and *NCSS* have been combined into one physical program. Access to each program is controlled by separate serial numbers. If you have a serial number for *PASS*, but not for *NCSS*, *NCSS* will work as a demo for 30 days from the time the first procedure is accessed.

If *NCSS* is already installed on your system, instruct the installation program to place this new version in the same folder as your previous version (usually \Program Files\NCSS97). All appropriate files will be replaced.

## What install does

The installation procedure (Setup) creates the necessary folders and copies the *PASS/NCSS* program from the installation file, called SETUP.EXE, to those folders. The files in SETUP.EXE are compressed, so the installation program decompresses these files as it copies them to your hard disk.

The folders created by Setup are (either NCSS97 or NCSS2000 may be substituted below):

\NCSS97 (or your substitute folder) contains most of the program files.

\NCSS97\DATA contains the database files used by the tutorials. We recommend creating a sub-folder of this folder to contain the data for each project you work on. For example, you might create a folder called \NCSS\DATA\Project1.

\NCSS97\JUNK contains temporary files used by the program while it is running. Under normal operation, *PASS* will automatically delete temporary files. After finishing *PASS*, you can delete any files left in this folder.

\NCSS97\REPORT is the default folder in which to save your output. You can save the reports to any folder you wish.

\NCSS97\SETTINGS contains the files used to store your template files. These files are used by the *PASS* template system which is described in a later chapter.

\NCSS97\STS contains all labels, text, and online messages.

\NCSS97\PDF contains printable copies of the documentation in Acrobat PDF format.

Setup places a file called NCSS97.INI in your windows folder. This file contains all default settings, paths, and constants that are used by the system. This file is documented in README.WRI. The settings in this file may be viewed and edited by selecting Edit then Options from the spreadsheet menu.

## Installing PASS and NCSS

This section gives instructions for installing *PASS* and *NCSS* on your computer system. You must use the *NCSS/PASS* setup program to install *PASS* and *NCSS*. The files are compressed, so you cannot simply copy the files to your hard drive.

After running the Setup program, you should read the README.WRI file for late-breaking information before starting the program.

Follow these basic steps to install *PASS* on your computer system.

### **Step    Notes**

1.     Make sure that you are using a 32-bit version of windows such as Windows 98, Windows Me, Windows NT 4.0, Windows 2000, or Windows XP.
2.     If you are installing from a CD, insert the CD in the CD drive. The installation program should start automatically. If it does not, on the Start menu, select the Run command. Enter 'D:Setup'. You may have to substitute the appropriate letter for your CD drive if it is not D. If you are installing from a download, simply run the downloaded file (SETUP.EXE).
3.     Once Setup starts, follow the instructions on the screen. *PASS* will be installed in the drive and folder you designate.

## If something goes wrong

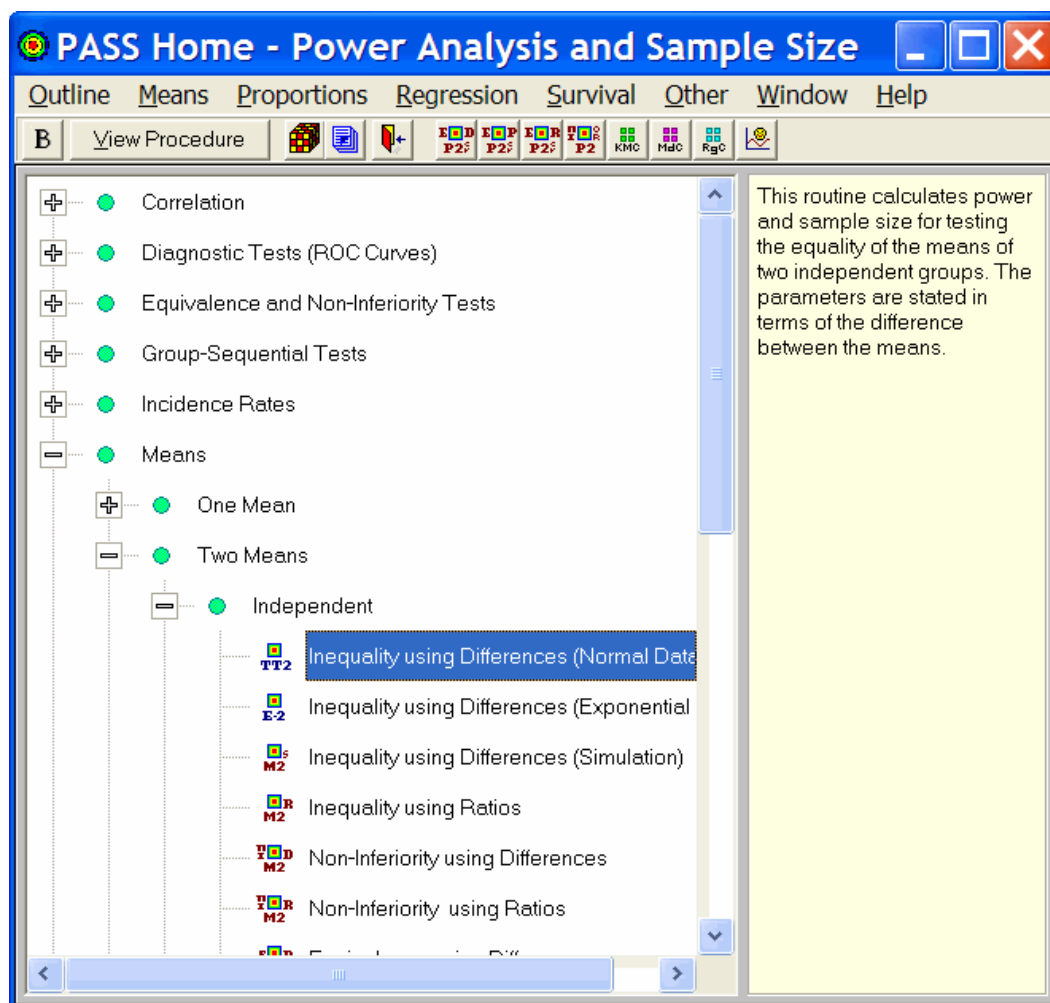
The installation procedure is automatic. If something goes wrong during installation, delete the \NCSS97 folder and start the installation process at the beginning. If trouble persists, contact our technical support staff as indicated below.



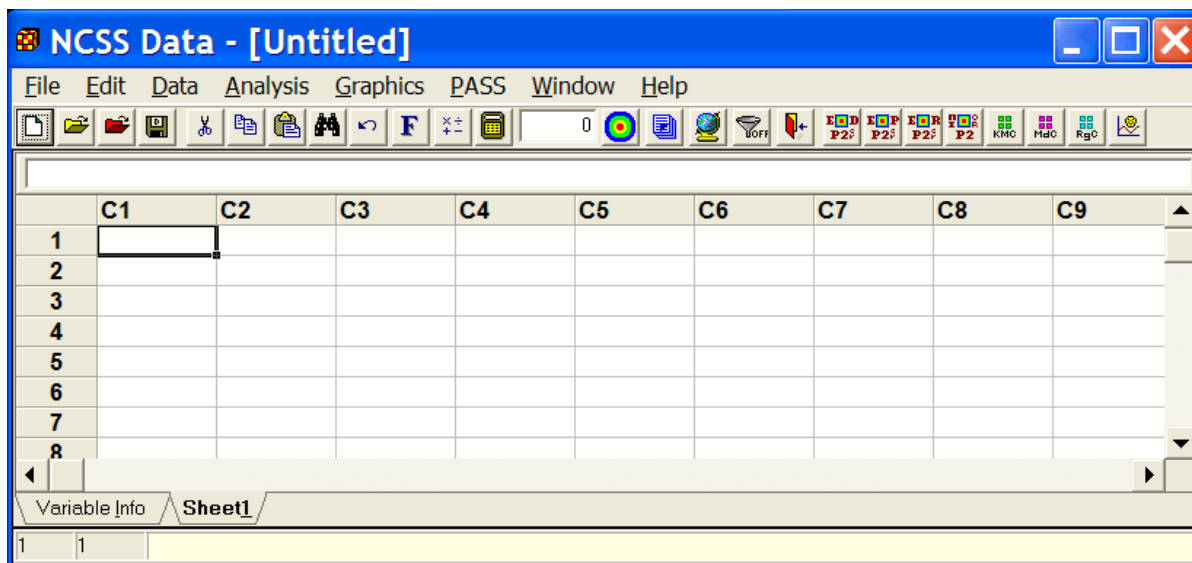
# Starting PASS

*PASS* may be started using your keyboard or your mouse using the same techniques that you use to start any other Windows application. You can start *PASS* by selecting *NCSS* from your Start menu using standard mouse or keyboard operations.

The first time you run *PASS*, enter your serial number in the pop-up window that appears when the program begins. If you have entered a serial number for *PASS* only (not for *NCSS*), the PASS Home window will appear.



If you have entered serial numbers for both *PASS* and *NCSS*, the *NCSS* spreadsheet window will appear.



To bring up the PASS Home window, click on the bull's-eye icon or select 'PASS Home' from the Window menu.

## Forcing the PASS Home screen to appear at startup

You can force the PASS Home window to appear when the program is run by taking the following steps:

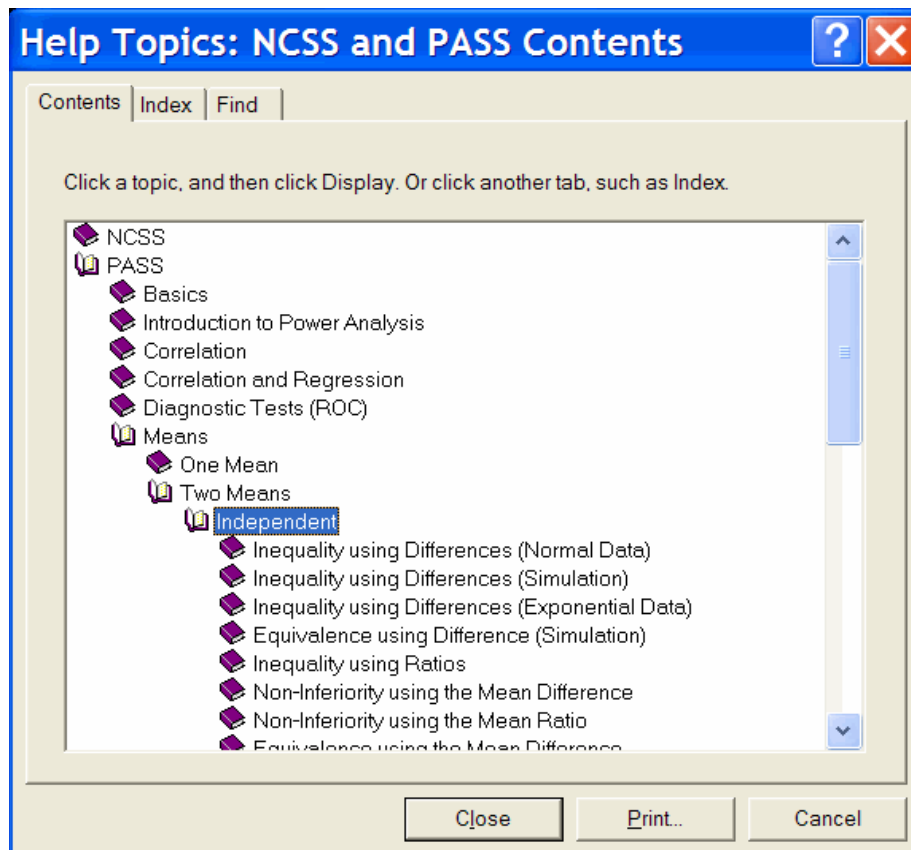
1. From the spreadsheet select Edit, then Options from the menus. Or, from the PASS Home window select Other, then Options from the menus.
2. Click on View tab.
3. Check the 'Show PASS Home Window' option.
4. Press the Ok button.

This will force the PASS Home window to appear first. You can view the spreadsheet at any time by clicking the yellow and orange dice icon or by selecting 'Data (Spreadsheet)' from the Window menu.

# Obtaining Help

## Online Help

To help you learn and use *PASS* efficiently, the material in the manuals is included in an online help system. To access this help system, select Help from the Help menu. When the help system is displayed, press the 'Contents' button at the top left of the window. This will display the following window with which you can browse the help system.



## Using Help

There are a few key features of our help system that, if you understand, will let you use the online help more efficiently. First, the Contents button brings up the table of contents of the help system. Use the Contents button to quickly navigate through the Help system. Second, each chapter was designed to be easily navigated. You can then proceed through a chapter section by section using the period and comma keys on your keyboard. Finally, you can use the Index and Find buttons to bring up an index of subjects.

## Printing Documentation

Obtaining a printed copy of the documentation is easy. Select 'View PDF File' from the Help menu. This will load and display the appropriate PDF file. From there, you can easily print a copy.

## Technical Support

To help us answer your questions more accurately, we may need to know about your computer system. Please have pertinent information about your computer and operating system available.

You can contact our technical support by calling **(801) 546-0445** between 8 a.m. and 5 p.m. (MST). You can contact us by email at **support@ncss.com**. Our goal is to respond to email within 24 hours.

## CHAPTER 2

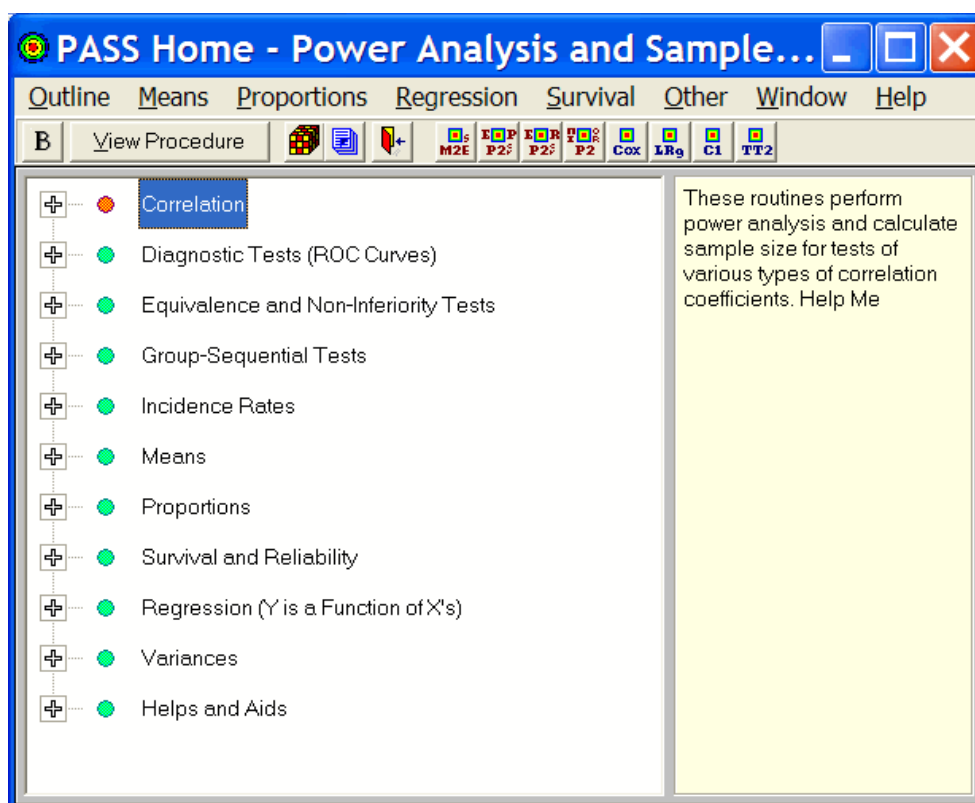
# Running PASS

## About this chapter

This chapter will show you how to start up and run a power analysis of the two-sample  $t$  test. It will give you a brief introduction to the windows used in *PASS*: the *PASS Home* window, the *procedure* window, and the *output* window.

## Starting *PASS*

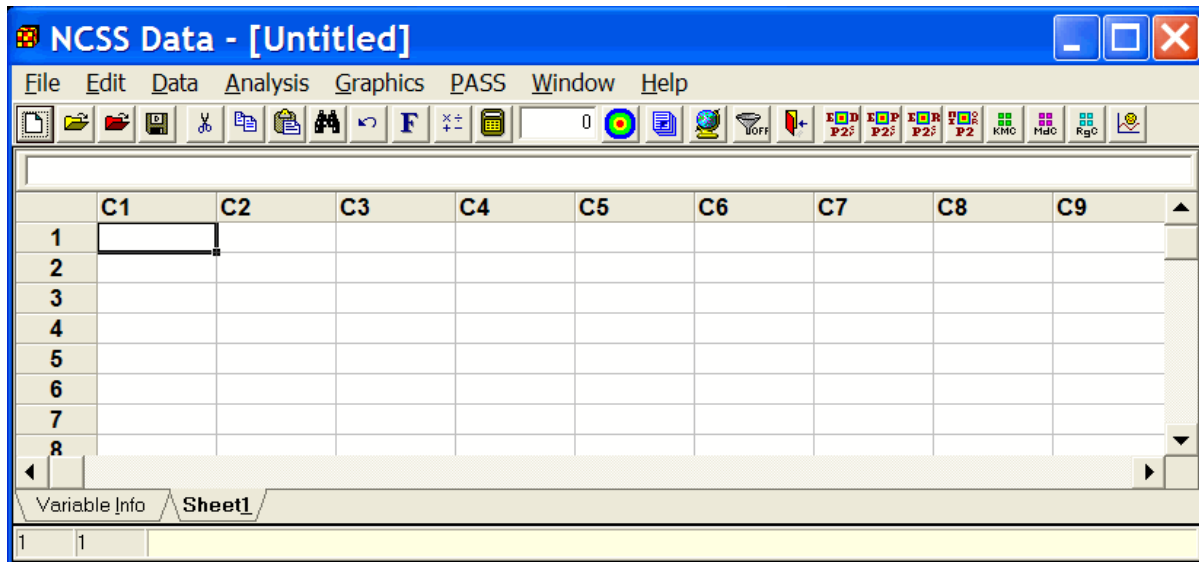
To start *PASS*, select *NCSS-PASS* from the Windows Start menu or double-click the *NCSS/PASS* icon. If you are licensed for *PASS*, but now for *NCSS*, the following *PASS Home* window will appear.



This window gives you access to all of the *PASS* procedures. Clicking on the plus sign or double-clicking on a phrase will expand the list so that you can see the procedures in that group. To load a specific procedure window, double-click on it or highlight it and click the View Procedure button.

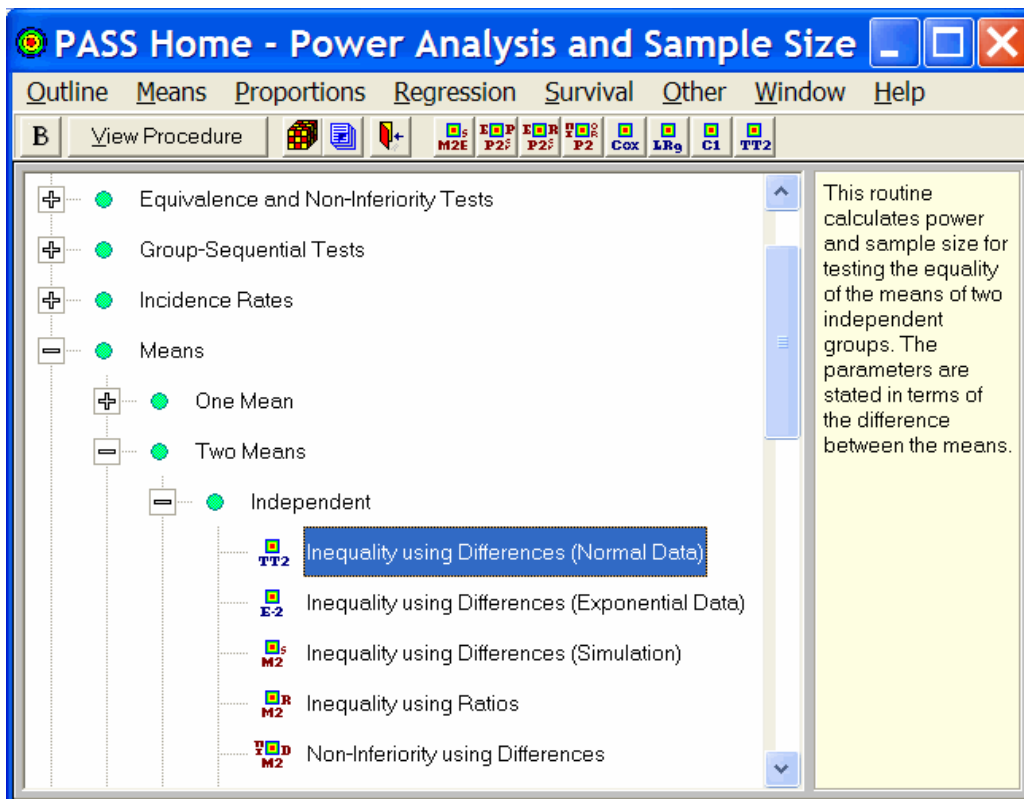
## 8 Running PASS

If you have previously installed *NCSS*, the *NCSS Data* window will appear first.



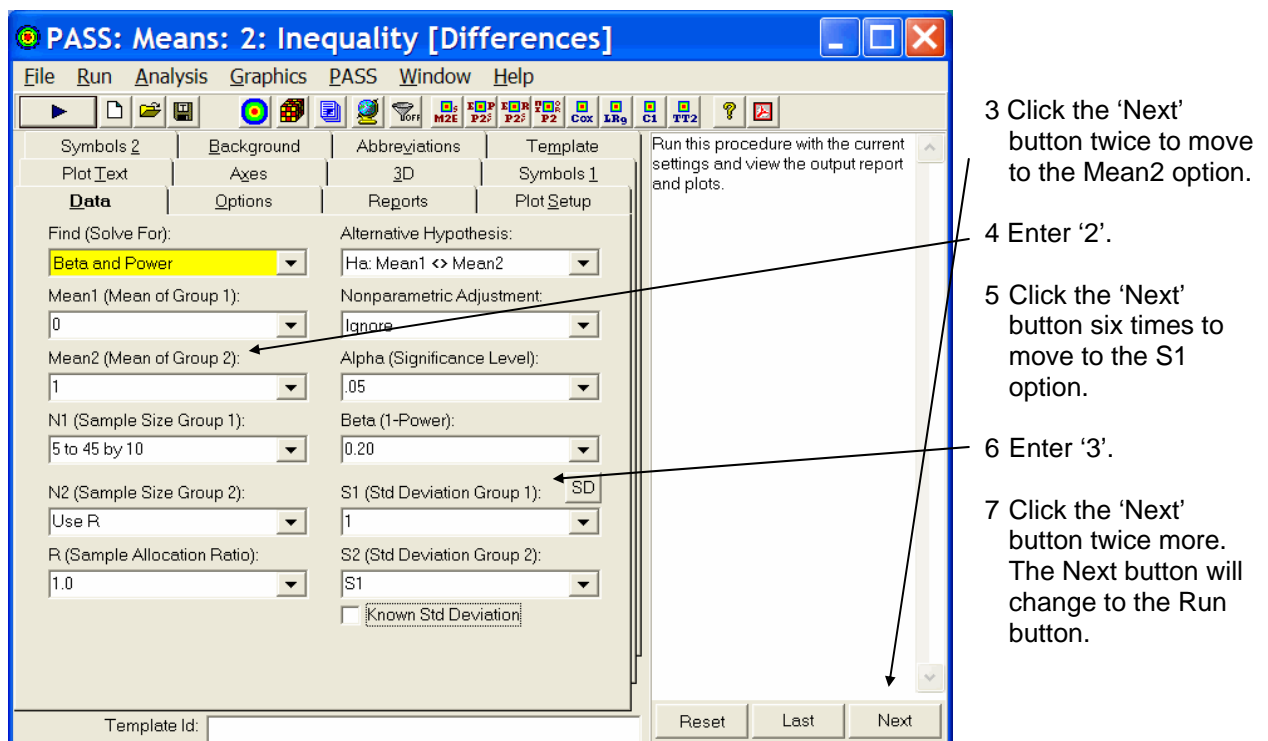
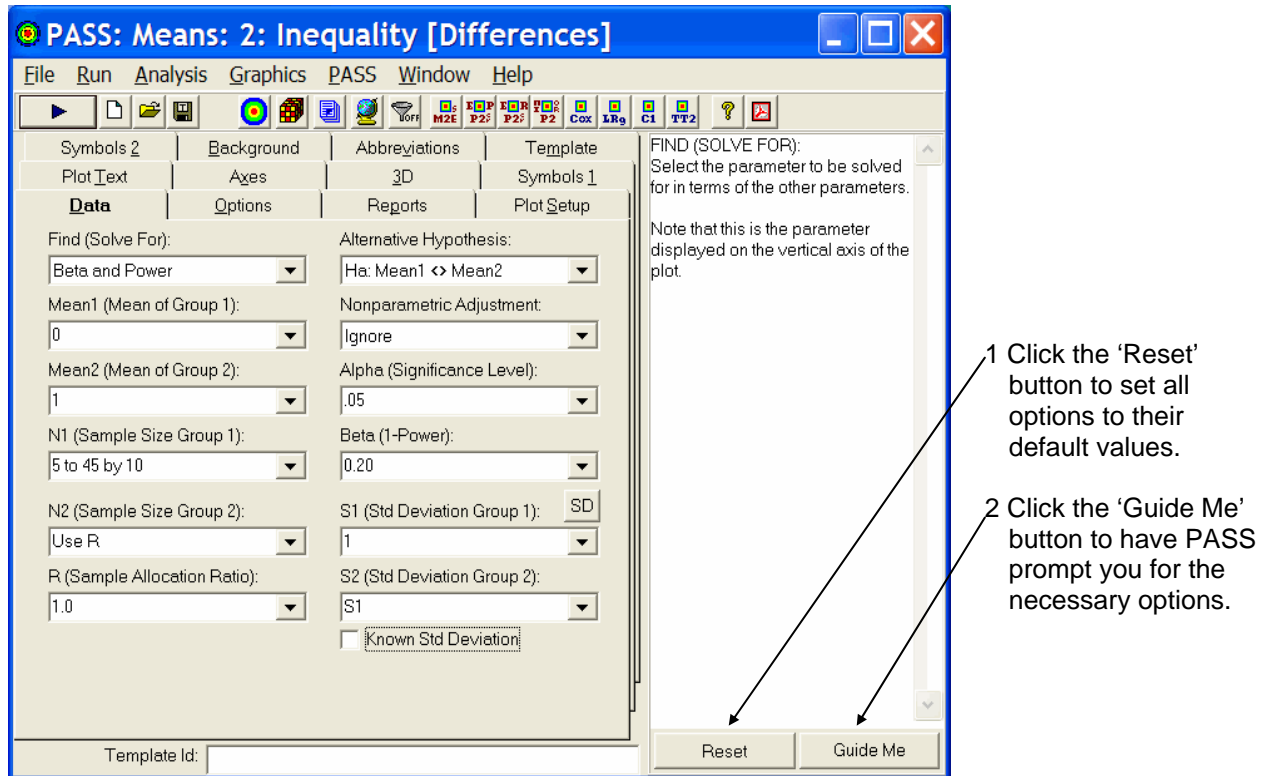
To bring up the PASS Home window, click on the bull's-eye icon or select 'PASS Home' from the Window menu.

The two-sample t-test is a procedure to test the inequality of two means from independent samples. Take the following steps to load this procedure. Expand the Means topic by double-clicking on the word 'Means'. Drilling down, double-click on 'Two Means', and then on 'Independent'. The first topic in the list is 'Inequality using Differences (Normal Data)'. This is the two-sample t-test. Double-click it.



The 'PASS: Means: 2: Inequality [Differences]' window will appear. Procedure windows let you specify, save, load, edit, and run an analysis.

We will run a power analysis using the default values except that the value of Mean2 will be 2 and the value of S1 will be 3.



## 10 Running PASS

The completed window will appear as follows.

**PASS: Means: 2: Inequality [Differences]**

File Run Analysis Graphics PASS Window Help

Symbols 2 Background Abbreviations Template  
Plot Text Axes 3D Symbols 1  
Data Options Reports Plot Setup

Find (Solve For): Beta and Power  
Mean1 (Mean of Group 1): 0  
Mean2 (Mean of Group 2): 2  
N1 (Sample Size Group 1): 5 to 45 by 10  
N2 (Sample Size Group 2): Use R  
R (Sample Allocation Ratio): 1.0  
Alternative Hypothesis: Ha: Mean1 <> Mean2  
Nonparametric Adjustment: Ignore  
Alpha (Significance Level): .05  
Beta (1-Power): 0.20  
S1 (Std Deviation Group 1): SD  
S2 (Std Deviation Group 2): S1  
☐ Known Std Deviation

Template Id: \_\_\_\_\_

Reset Last Run

S2 (Std Deviation):  
Enter an estimate of the standard deviation of group 2. The standard deviation must be a positive number.  
Enter S1 if you want to use the same value(s) as those for group 1.  
Press the 'SD' button to obtain help on estimating the standard deviation.  
You can enter a range of values such as 1,2,3 or 1 to 10 by 1.

8 Click the 'Run' button to calculate the power analysis and display the following report.

**NCSS Output - [PASS: Means: 2: Inequality [Differences] Output]**

File Edit View Format Window Help

Two-Sample T-Test Power Analysis

Page/Date/Time 1 10:22:02 AM

**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	Ratio	Alpha	Beta	Mean1	Mean2	S1	S2
0.15387	5	5	1.000	0.05000	0.84613	0.0	2.0	3.0	3.0
0.42206	15	15	1.000	0.05000	0.57794	0.0	2.0	3.0	3.0
0.63673	25	25	1.000	0.05000	0.36327	0.0	2.0	3.0	3.0
0.78504	35	35	1.000	0.05000	0.21496	0.0	2.0	3.0	3.0
0.87852	45	45	1.000	0.05000	0.12148	0.0	2.0	3.0	3.0

**References**  
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, MA.  
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

**Report Definitions**  
Power is the 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.

Page 1/2 Line 2 Col 21

The **Output** window displays the output of the power analysis. It serves as a mini-word processor—allowing you to view, edit, save, and print your output.

You may want to scroll down to view the graph at the end of the report.

When you are finished, you can quit **PASS** by selecting *Exit* from the *File* menu.



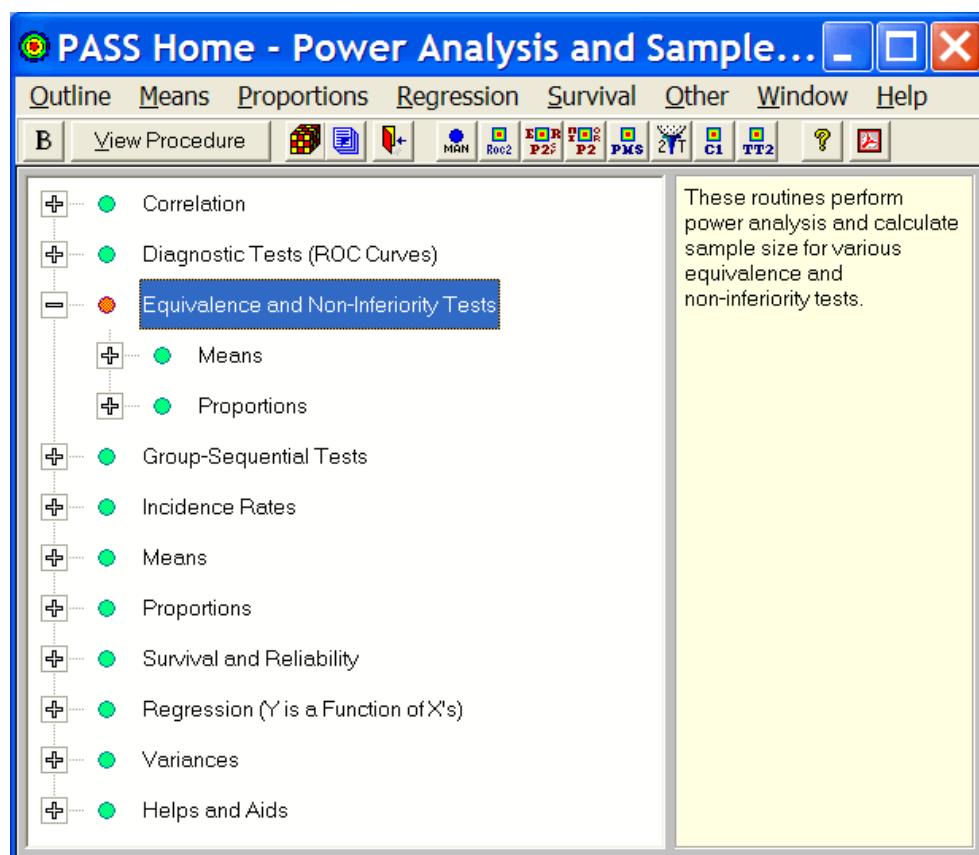
## CHAPTER 3

# PASS Home Window


## Introduction

The *PASS Home* window lets you quickly and easily find the appropriate procedure to be loaded. Using an outline format, it lists every procedure in *PASS* along with a brief statement that describes what the procedure is for and when it might be used.

The PASS Home window also lets you configure the eight procedure buttons that appear on the toolbars of the Data, Output, and Procedure windows. These buttons give you immediate access to your favorite procedures.



## Using the PASS Home

The PASS Home window is easy to use. It is loaded automatically if you are registered for *PASS*, but not *NCSS*. If you are registered for *NCSS*, press the bulls-eye icon  center top of the spreadsheet to load the PASS Home window. Alternatively, you can select the PASS Home option from the Window menu.

The PASS Home window has a set of menus, a toolbar, and a large display area. On the left side of the display area is an outline list of all the procedures in *PASS*. On the right side of the display area is the immediate help area that displays a brief statement explaining the currently selected item to the left.

## Menus

### Outline Menu

#### Collapse Outline

This option collapses the outline so that only the main heading is displayed.

#### Expand to First Level

This option expands the outline so that the main headings and first-level subheadings are displayed.

#### Expand All

This option completely expands the outline so that all entries are displayed.

#### Bold Text

This option toggles the bolding of the text.

#### Goto Selected Procedure

This option loads the currently selected procedure's window.

#### Close

This option closes the PASS Home window.

### Window Menu

This menu allows you to open other windows in the *PASS/NCSS* system such as the Data window or the Output window.

### Help Menu


This menu allows you to view the help system, modify your serial numbers, and load various portions of the printable (PDF) documentation.

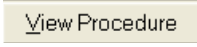
# Toolbar


The toolbar gives you one-click access to several of the menu items. The menu item assigned to each button on the toolbar is displayed when the mouse is held over the button for a few seconds.




The action caused by each of these icons is discussed next.


 This icon toggles the bolding of the text in the outline window.


 This icon causes the window of the currently selected procedure to be displayed. You can accomplish the same action by double-clicking on the procedure name.


 This icon causes the NCSS spreadsheet window to be displayed. You will find this useful when working with procedures that use the spreadsheet, such as the Repeated Measures ANOVA procedure.

 This icon causes the output window to be displayed.

 This icon closes all windows and exits the program.

 These buttons show up on all toolbars throughout the *NCSS/PASS* system. Clicking on them with the left mouse button will display that procedure. Clicking on one of these buttons with the right mouse button changes the button to the highlighted procedure.

 This icon loads the help system.

 This icon loads the file containing this chapter in PDF format.

## Customizing the Toolbars

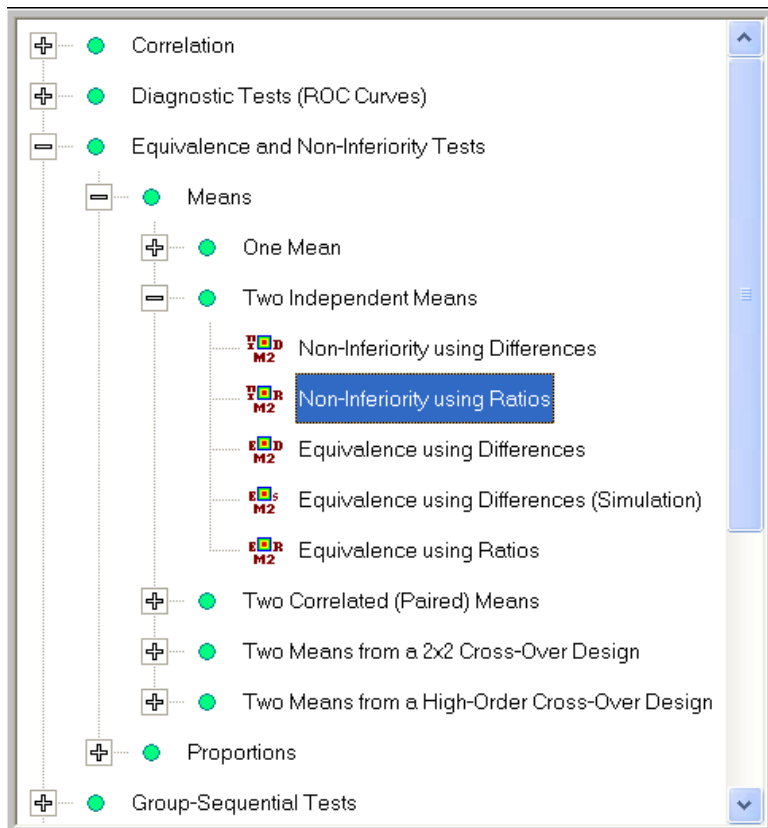
The eight procedure buttons that show up on all toolbars throughout the program may be changed here. The process of assigning one of these eight buttons a new procedure is as follows:

- 1 Find and select the procedure in the outline section (left-side of main window) of the PASS Home window.
- 2 Click on the button you want to assign the procedure to with the right-mouse button.

That's it. The icon of the selected procedure will now appear in all toolbars throughout the program.

# Outline

The outline expands and contracts as you either click on a plus or minus sign, or double-click on a topic. This gives you quick, intuitive access to all of the procedures in *PASS*.



In the example shown here, we clicked on “Equivalence and Non-Inferiority Tests”, then on “Means”, then on “Two Independent Means”, and finally on “Non-Inferiority using Ratios” to highlight it. If we double-clicked on “Non-Inferiority using Ratios”, that procedure would be displayed.

## CHAPTER 4

# Procedure Window

## Introduction

All *PASS* procedures are controlled by a *procedure* window that contains all of the settings, options, and parameters that control the input and output of the program. These options are separated into groups called *panels*. A particular panel is viewed by pressing the corresponding tab that appears near the top of the window. For example, in the window below, the Data panel is active. Other panels are Options, Reports, and Plot Setup.

The values of all options available for a procedure are referred to as a *template*. A template may be stored for future use in a *template file*. By creating and saving template files (often referred to as *templates*), you can tailor each procedure to your own specific needs. For example, you may want to your plots to be bar charts. This capability can be saved as a template. Each time you use a procedure, you simply load your template and run the analysis you have preset. You do not have to set all the options every time. The specific operations needed to do these are shown later.

Note that at most six procedure windows can be opened at a time. Also note that you can widen the window to increase the size of the immediate help window by dragging the corners of the window.

The screenshot shows the **PASS: Means: 2: Inequality [Differences]** window. The window has a menu bar (File, Run, Analysis, Graphics, PASS, Window, Help) and a toolbar with various icons. Below the toolbar are several tabs: Symbols 2, Background, Abbreviations, Template, Plot Text, Axes, 3D, Symbols 1, Data (active), Options, Reports, and Plot Setup. The **Data** panel contains the following settings:

- Find (Solve For):** Beta and Power
- Mean1 (Mean of Group 1):** 0
- Mean2 (Mean of Group 2):** 1
- N1 (Sample Size Group 1):** 5 to 45 by 10
- N2 (Sample Size Group 2):** Use R
- R (Sample Allocation Ratio):** 1.0
- Alternative Hypothesis:** Ha: Mean1 <> Mean2
- Nonparametric Adjustment:** Ignore
- Alpha (Significance Level):** .05
- Beta (1-Power):** 0.20
- S1 (Std Deviation Group 1):** SD
- S2 (Std Deviation Group 2):** S1
- ☐ Known Std Deviation

On the right side of the window, there is a **FIND (SOLVE FOR):** section with the text: "Select the parameter to be solved for in terms of the other parameters. Note that this is the parameter displayed on the vertical axis of the plot." At the bottom of the window, there is a **Template Id:** field and two buttons: **Reset** and **Guide Me**.

## Default Template

Whenever you close a procedure, the current settings are automatically saved in a default template file named *default*. This template file is automatically loaded when the procedure is next opened. This allows you to continue using the template without resetting all of the options.

## Procedure Window Anatomy

This section explains the various objects found on the template.

**Menus.** The menus let you move to other windows.

**Run.** Clicking this button runs the program and generates output.

**Options.** These fields set values that control the analysis.

**Template Id.** This box can contain a phrase that identifies this template.

**Reset.** This button resets all options under all tabs to their default values.

**Guide Me.** This button instructs the program to step you through the main options that must be set for an analysis.

**Tabs.** The tabs let you view different groups of options.

**Immediate Help.** This box displays a brief help message about the field that the mouse is currently positioned over.

The screenshot shows the PASS: Means: 2: Inequality [Differences] window. It features a menu bar (File, Run, Analysis, Graphics, PASS, Window, Help) and a toolbar with icons for various functions. The main area is divided into tabs: Symbols 2, Background, Abbreviations, Template, Symbols 1, Data, Options, Reports, and Plot Setup. The 'Data' tab is currently selected, showing fields for 'Find (Solve For):', 'Mean1 (Mean of Group 1):', 'Mean2 (Mean of Group 2):', 'N1 (Sample Size Group 1):', 'N2 (Sample Size Group 2):', 'R (Sample Allocation Ratio):', 'Alternative Hypothesis:', 'Nonparametric Adjustment:', 'Alpha (Significance Level):', 'Beta (1-Power):', 'S1 (Std Deviation Group 1):', 'S2 (Std Deviation Group 2):', and a 'Known Std Deviation' checkbox. A 'Template Id:' field is at the bottom left, and 'Reset' and 'Guide Me' buttons are at the bottom right. An 'Immediate Help' box on the right displays a message about the 'Find (Solve For):' field.

# Menus

The menus provide a convenient way to transfer from module to module within the *PASS* system. Each set of menus will be briefly described here.

## File Menu

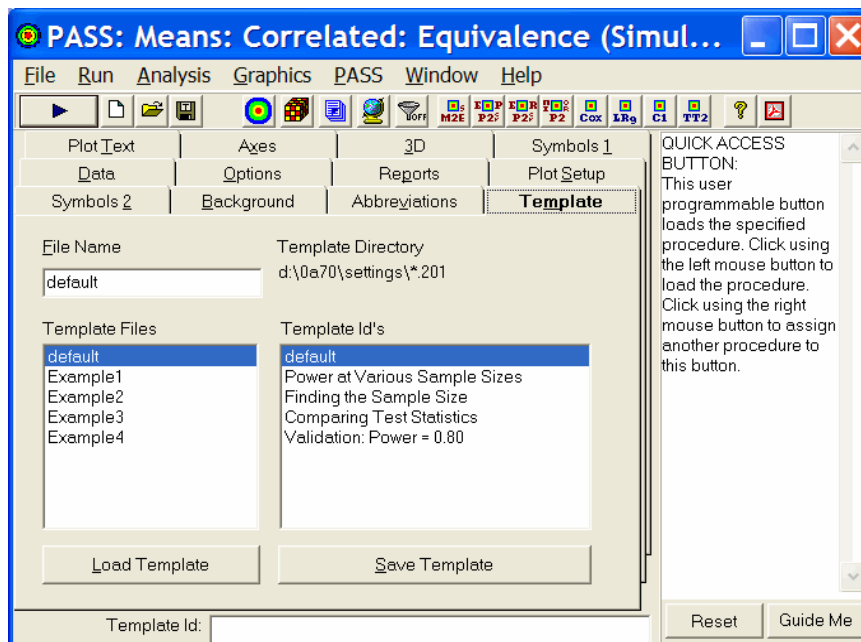
The File Menu is used for initializing, loading, and saving a copy of a template. Each set of options for a procedure, called a *template*, may be saved for future use. In this way, you do not have to set the options every time you use a procedure. Instead, you set the options the first time, save them as a template, and re-use the template whenever you re-use the procedure.

### New Template (Reset)

This menu item resets all options to their default values. It performs the same function as the Reset buttons.

### Open Template Panel

This option sets the Template panel as the active procedure panel. The Template panel lets you load or save template files. It displays all templates associated with this procedure along with the Template Id (the optional phrase at the bottom of the window).



### Save Template (button)

To save a template, enter the name you want to give the template file in the File Name box. You may also enter an identifying phrase in the box at the bottom of the window since this will be displayed along side of the file names. Finally, press the Save Template button to save the file.

Note that there is no automatic connection between the template in memory and the copy on

the disk. If you want to save the changes you have made to a template, you must use the Save Template option to save them.

### Load Template (button)

To load a template file, select it from the list of files given in the Template Files box. Once the desired file is selected, press the Load Template button to load the template.

### Save Template

This option saves the current option settings to the template file that is currently specified in the File Name option of the Template panel. You can be viewing any panel of the procedure when you issue this command—you do not have to be viewing the Template panel.

The template files are stored in the Settings folder. You can erase any unwanted template files by deleting them from this folder using the Windows Explorer program.

The template files for each procedure have different file name extensions. Thus, you can use the same name for a template saved from the T-Test procedure as for a template saved from the Multiple Regression procedure. For example, if the ‘Save Template’ command is issued in the window shown above, the current settings will be saved in a file called ‘default.201’ in the Settings folder.

The Save button on the toolbar provides this same operation. It may be more convenient than selecting this menu item.

### Close Procedure

This option closes this procedure window.

### Printer Setup

This option lets you set various printer options.

### Exit NCSS and PASS

This option terminates the *NCSS/PASS* system. Before using this option, you should save all datasheets, templates, and output documents that you want to keep.

### Run Menu

This menu controls the execution of the program.

### Run Procedure

The Run Procedure option runs the analysis, displaying the output in the Output document of the word processor. After you have set all options to their appropriate values, select this option to perform the analysis.

Note that the procedure may also be run by pressing the *F9* function key or by pressing the left-most key on the toolbar (the dark-blue-arrow button).

### Abort

After starting a procedure, you may find that it is taking longer than you anticipated to finish. You can stop the running of the procedure by pressing this button. The red stop-sign icon that appears on the top right of the screen may be pressed for the same purpose.

### Analysis and Graphics Menus

These menus allow you to transfer to various *NCSS* procedures. You can load these procedure windows, but when you try to run them, you will receive an error message indicating that you do not have a license for this procedure (unless you have purchased a license for *NCSS*).



## PASS Menu

This menu allows you to directly load any of the various *PASS* procedures.

## Window Menu

This menu lets you display any of the other windows in the *NCSS/PASS* system that are currently open such as the Output window, the Data (Spreadsheet) window, the Navigator window, or any procedure windows.

### Output

Select this option to display the output window.

### Data (Spreadsheet)

Select this option to display the *NCSS* spreadsheet.

### Navigator

Select this option to display the *NCSS* navigator window.

### Reset Window Positions

Occasionally, *NCSS/PASS* windows will be loaded, but will not display. This menu item will load the Options window to a tab that will let you reset the position of all program windows.

## Help Menu

This menu gives you access to the *PASS* documentation.

### Help...

This option loads the help system. Once loaded, press the Contents button to obtain a table of contents window.

### About...

This option show you which products are licensed and when your version was released.

### Serial Numbers...

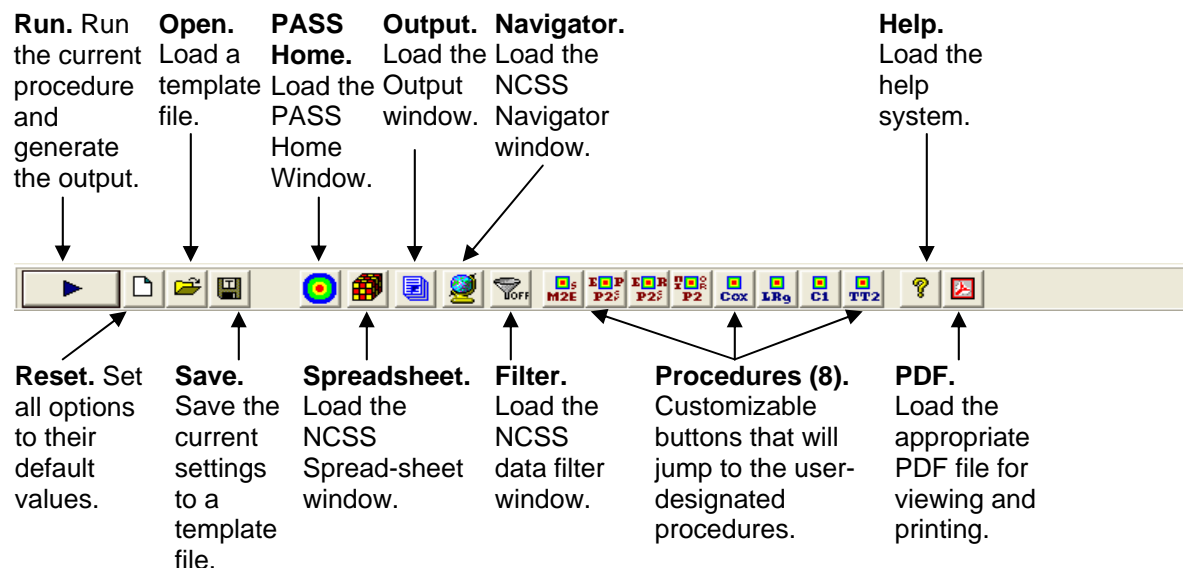
This option loads the serial number screen. Use this when you need to change one of your serial numbers.

### View PDF File...

This option loads the PDF file that documents this procedure. This file may then be printed.

# Toolbar

The toolbar is a series of small buttons that appear just below the menus at the top of the procedure window. Each of these buttons provides quick access to a menu item.



## Important Tip for PASS Users

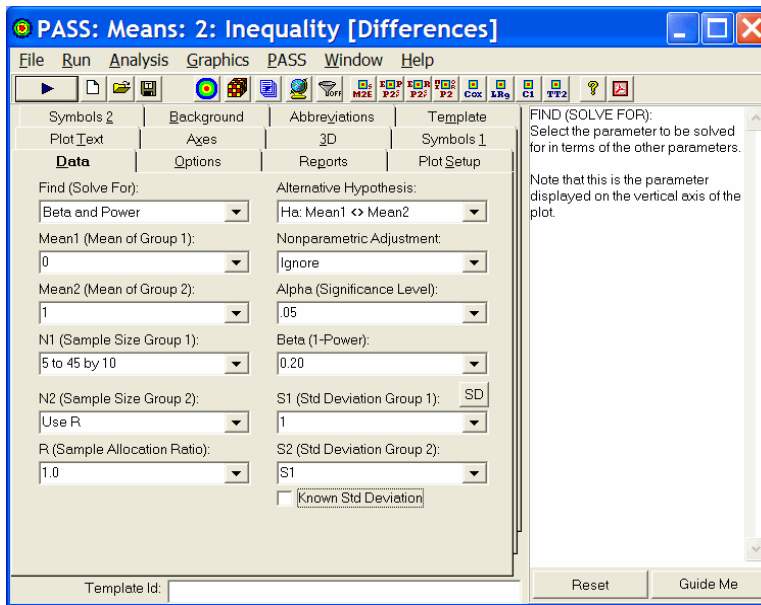
The Procedures (8) buttons are a series of eight buttons that can be changed to your favorite procedures by right clicking on any of these buttons.

# The Panel Tabs

The procedure window contains several sets of options (panels). Each panel is displayed by clicking on the appropriate tab. We will now describe the purpose and operation of each panel.

## Data tab

This tab displays most of the options specific to the procedure. This is where you set the values of power, sample size, alpha, etc. These options are described in detail in the chapters corresponding to each procedure. Once you have set the options, click the **Run** button to generate the output.



## Entering Multiple Values

In most cases, boxes that are extra wide allow you to enter multiple values. When this is done, a separate analysis is made for each combination of all multiple values. For example, if you enter four sample sizes and three alpha values, the resulting report will contain  $3 \times 4 = 12$  rows, one for each combination.

You can enter multiple options using list or the *to-by* syntax. The *to-by* syntax is most easily described by an example.

The *to-by* phrase *20 to 100 by 20*

is translated to the values: *20,40,60,80,100*.

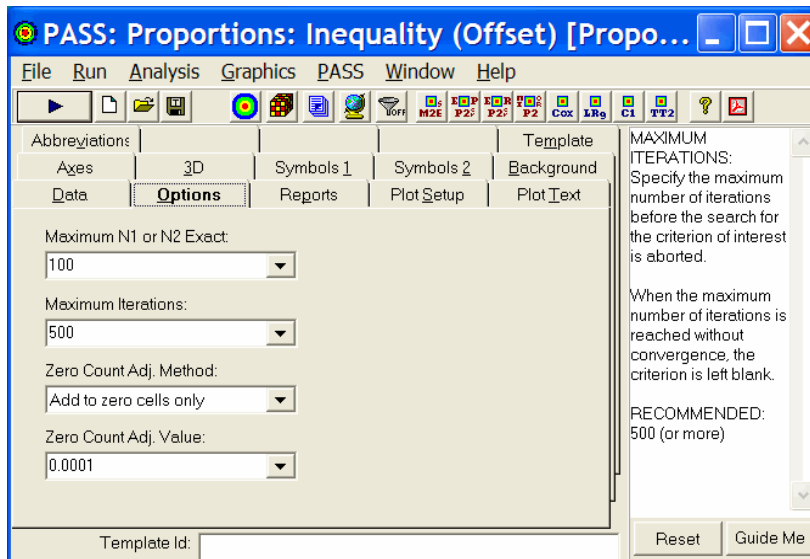
## Find (Solve For)

Specify the parameter that is to be solved for in terms of the other parameters. For example, you might want to solve for power or sample size.

In most cases, the algorithm for the calculating the power is programmed within **PASS**. When other parameters (such as sample size or difference) are selected, a binary search is conducted using the power algorithm. These searches can be time consuming, so the best place to start is with this option set to 'Beta and Power'.

## Options tab

The Options window presents the parameters that control the searching process.

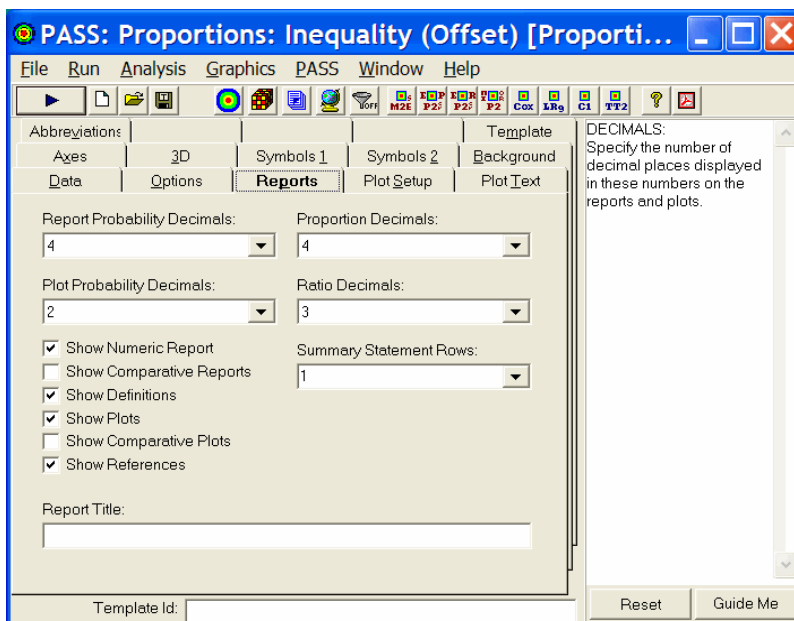


### Maximum Iterations.

This option specifies the maximum number of iterations before a search for the parameter of interest is halted. When the maximum number of iterations is reached without convergence, the criterion is left blank. We recommend that at least 500 iterations be specified.

## Reports tab

This tab displays the options that control the output reports.



### Decimals

These options set the number of decimal places in corresponding values of the numeric and graphic output.

### Summary Statement Rows

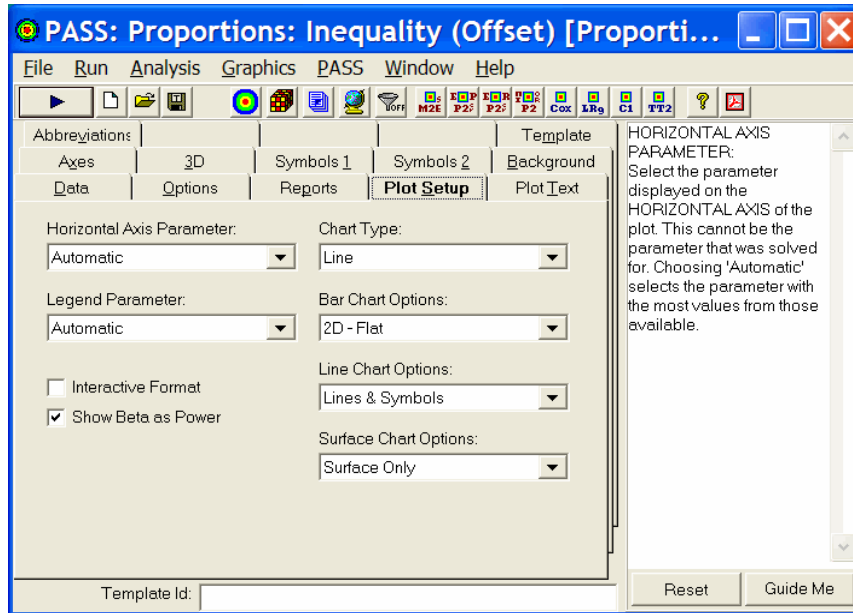
The Summary Statement is a paragraph that can be output for each row of the report. When you do not need this output, you can reduce the number of rows accordingly. To eliminate the text output completely, set this value to zero.

### Show ...

These options control whether the corresponding item is displayed in the output.

## Plot Setup tab

This tab displays options that control the plots, including which parameters are shown on the axes and the type of plot that is displayed.



### Horizontal Axis Parameter

This option selects which of the parameters from the Data tab is displayed across the horizontal axis. Note that you cannot select the parameter that was listed in the Find option. Also, you would normally only select a parameter that has multiple entries.

When this option is set to *Automatic*, the parameter with the most values is selected.

### Legend Parameter

A separate line is drawn for each value of this parameter. The lines are labeled in the legend. When this option is set to *Automatic*, the parameter with the second most values is selected.

### Show Beta as Power

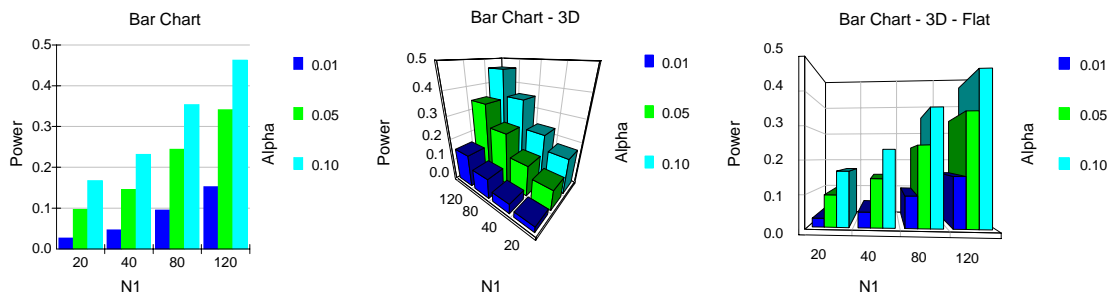
This option controls whether Beta or Power is displayed on the plots.

### Chart Type

This option controls the type of chart that is displayed. Bar charts, line charts, and surface charts are available.

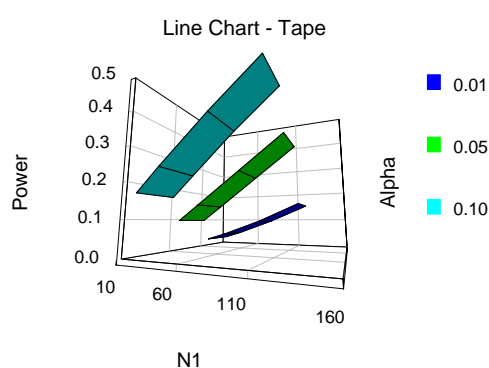
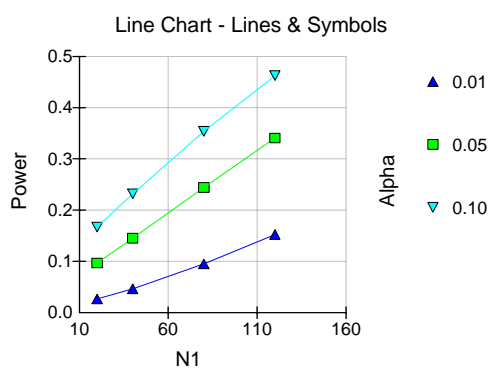
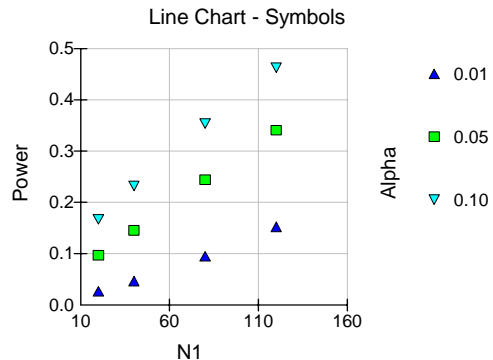
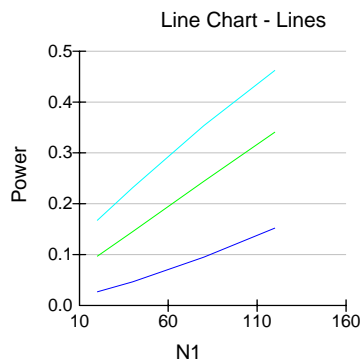
### Bar Chart Options

These chart types are available when Chart Type is set to Bar.



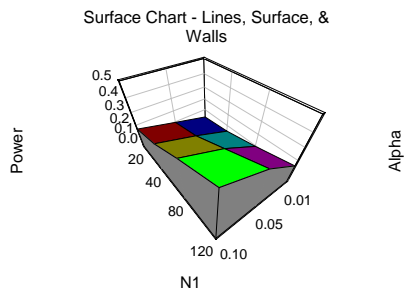
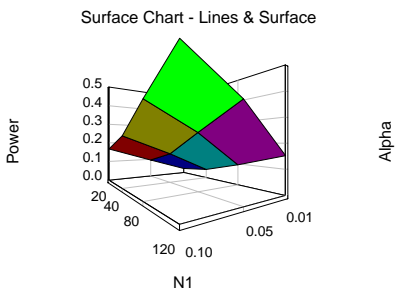
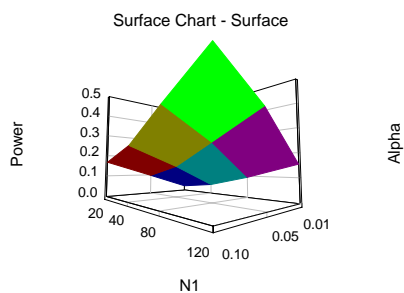
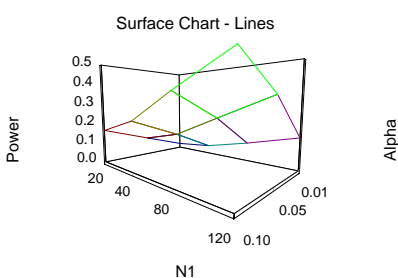
## Line Chart Options

These chart types are available when Chart Type is set to Line.



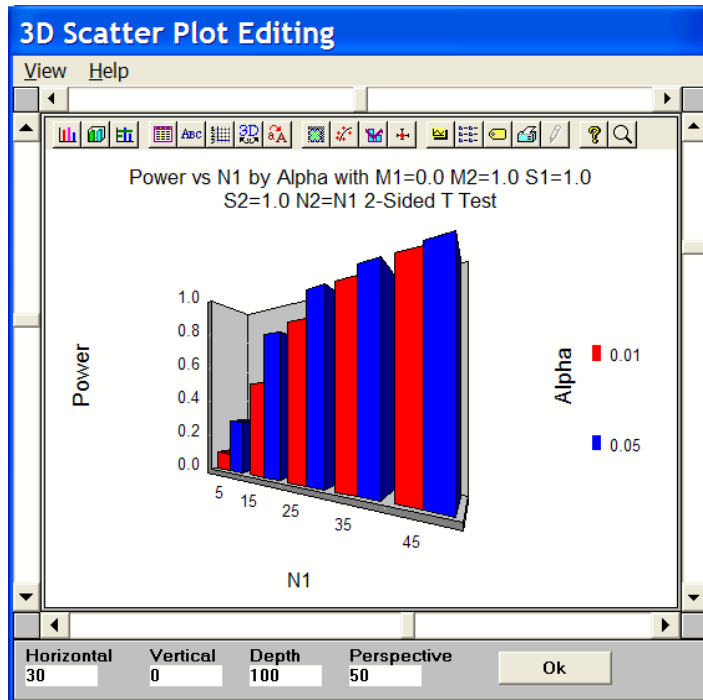
## Surface Charts

These chart types are available when Chart Type is set to Surface.



## Interactive Format

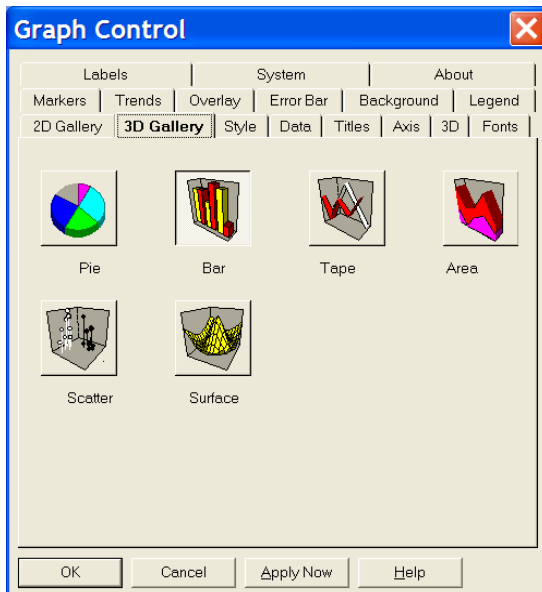
This option controls whether the plot may be reformatted interactively after it has been generated. When checked, this option allows charts to be formatted interactively using the following window.



The four scroll bars around the edge of this window control the vertical axis, horizontal axis, depth, and perspective. The current values of these parameters are shown in the boxes at the bottom of the screen.

Once you are finished editing chart, click the *Ok* button to proceed.

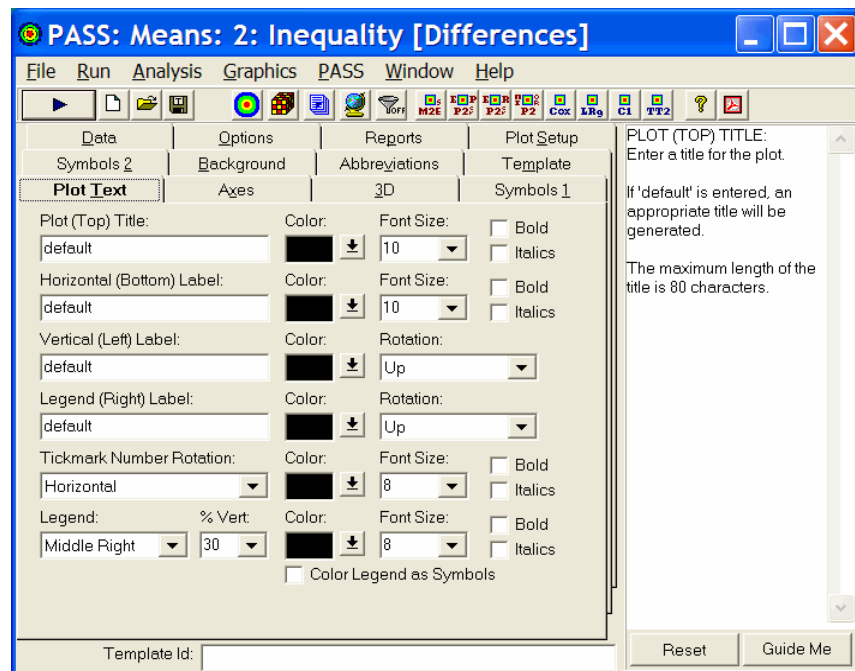
Each of the buttons along the top of the Scatter Plot Editing Window will display a different tab of the Graph Control window. Each tab provides options which allow detailed modification of the chart.



We will not document these options here since most of them are not necessary to the running of *PASS*. If you want to explore these options further, choose the Help button at the bottom of the window. This will bring up a special help system that describes all graphics options in detail.

## Plot Text tab

This window controls the titles and labels of the plots.



### Plot Title - Legend Label

The title and label options specify the text, color, and font size of the corresponding value displayed on the plot.

### Tickmark Number Rotation

This option specifies whether the reference values are shown vertical or horizontal.

### Legend

This option sets the position of the legend

### % Vert

Specify the size of the legend area as a percentage of the maximum possible. This option lets you shrink a legend that is too large

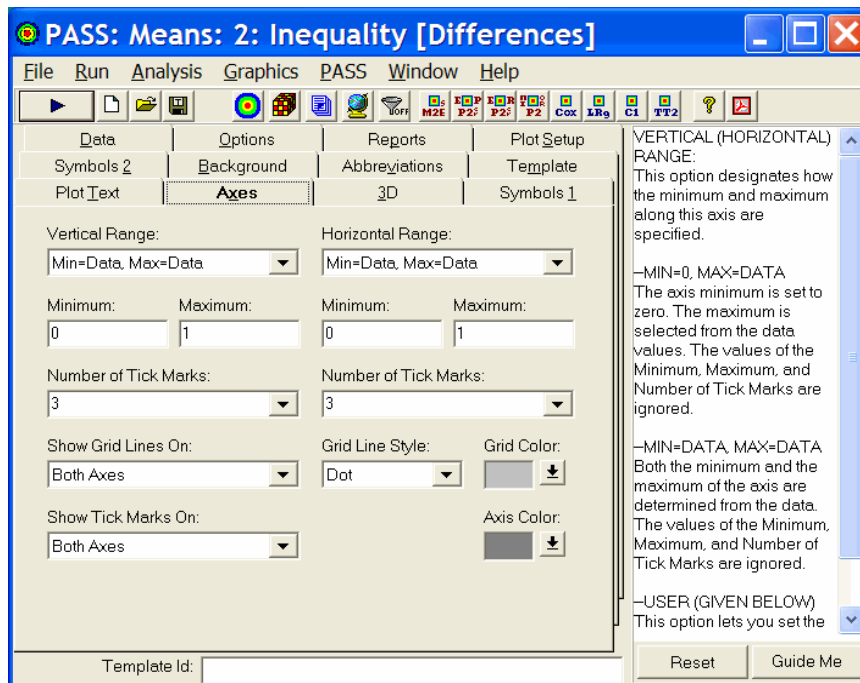
### Color Legend as Symbols

Normally, text in the legend is displayed using the color selected by the Color option. This option indicates that each legend entry is to be displayed in the corresponding group color.



## Axes tab

This tab displays options that control the chart axes.



## Vertical / Horizontal Range

This option designates how the minimum and maximum along this axis are specified. Available options are:

**Min=0, Max=Data.** The axis minimum is set to zero. The maximum is selected from the data values. The values of the Minimum, Maximum, and Number of Tick Marks are ignored.

**Min=Data, Max=Data.** Both the minimum and the maximum of the axis are determined from the data. The values of the Minimum, Maximum, and Number of Tick Marks are ignored.

**User.** This option lets you set the Minimum, Maximum, and Number of Tick Marks to scale the axis. These options determine which of the axes have grid lines displayed. This option is particularly useful when you want to make sure that the axis displaying power values displays a grid between zero and one.

## Minimum and Maximum.

Specify the axis minimum or maximum to be used when the Vertical (or Horizontal) Range option is set to User.

## Number of Tick Marks

Specify the number of tick marks along this axis. This value is used when Vertical (or Horizontal) Range is set to User.

## Show Grid Lines On

These options determine which of the axes have grid lines displayed.

## Grid Line Style

Specify the pattern of the line. This option only works when the line width is 20.

## Grid Color

Specify the color of the grid line.

## Show Tick Marks On

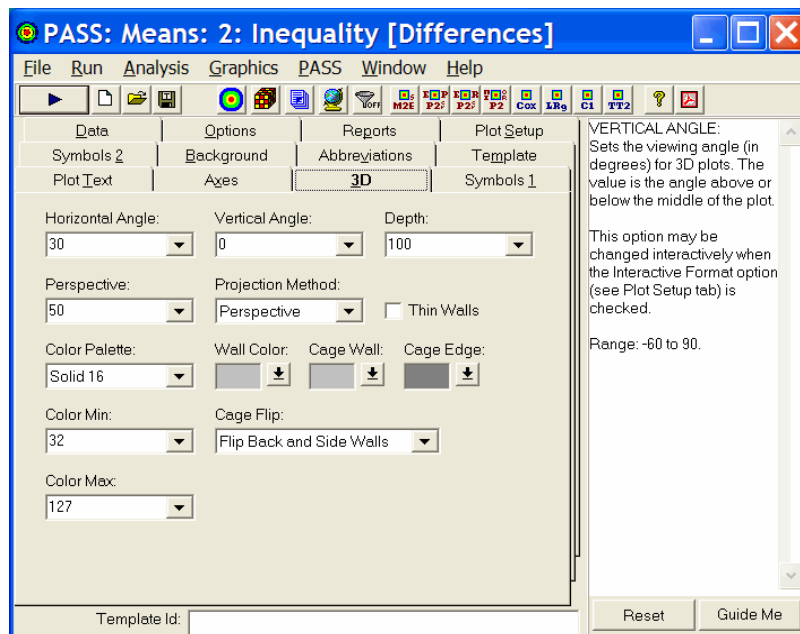
These options control which of the axes have tick marks displayed.

## Axis Color

Specify the color of the Axis line.

## 3D tab

This tab displays options that control the 3D charts.



## Horizontal Angle

This option sets the horizontal viewing angle (in degrees) for 3D plots. It represents an angle around the base of the plot.

The range of values is -180 to 180 degrees. This option may be changed interactively when the Interactive Format option is checked.

## Vertical Angle

This option sets the vertical viewing angle (in degrees) for 3D plots. It represents an angle above or below a point halfway up the graph. Values may range from -60 to 90 degrees.

This option may be changed interactively when the Interactive Format option is checked.

## Depth

This option sets the projected depth of 3D plots. Depth is a percentage of 100, calculated to provide equal increments in the X and Z directions. Values may range from 5 to 400. This option may be changed interactively when the Interactive Format option is checked.

## Perspective

This option sets the degree of perspective foreshortening in 3D plots. Perspective is the perceived distance of the viewer from the graph. The range of values is 0 to 100. This option may be changed interactively when the Interactive Format option is checked.

## Projection Method

Sets the projection method of 3D-type charts.

## Off

No graph is drawn.

**Isometric**

The graph is drawn, but no perspective is attempted.

**Perspective**

The axes are tilted to give a 3D perspective to the plot.

**Thin Walls**

This option specifies whether the walls of the axis grid that form the background of the chart are thick or thin.

**Color Palette**

Specify a color palette for the surface chart. Using a setting of, say, Black to Red will allow the surface plot to show a continuous array of red hues from lowest to highest.

**Wall Color**

Specify the color of the wall.

**Cage Wall**

Specify the color of the cage (grid) wall.

**Cage Edge**

Specify the color of the cage (grid) edge.

**Color Min**

Specifies the number of the color to be associated with the lowest numerical value. Possible values are 32 to 127. A value near 50 usually works well. Note that this option only works with 128-color palettes.

**Color Max**

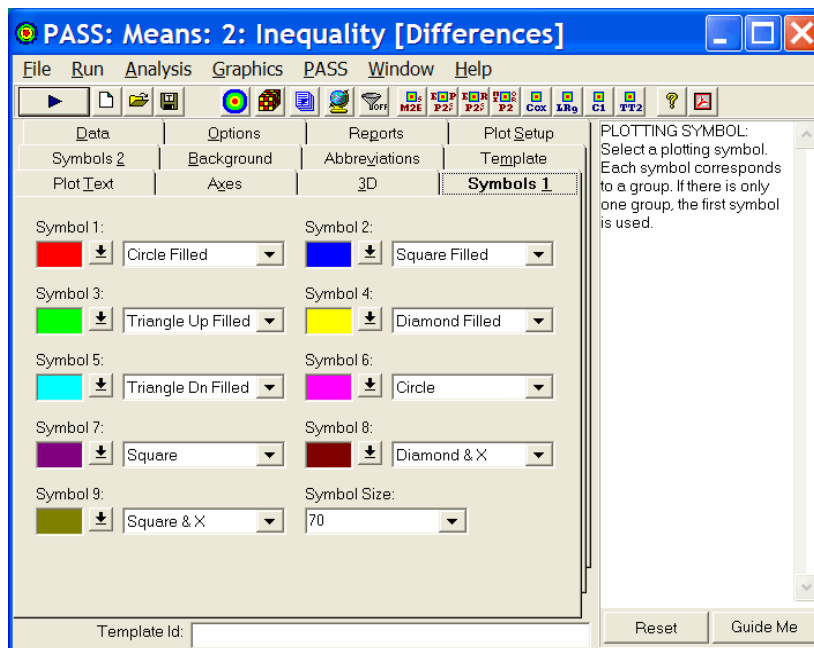
Specifies the number of the color to be associated with the largest numerical value. Possible values are 32 to 127. A value near 120 usually works well. Note that this option only works with 128-color palettes.

**Cage Flip**

This option controls whether the back and side walls of the graph cage are allowed to switch to the opposite edge for better viewing as the graph is rotated.

## Symbols 1 and 2 tabs

These tabs specify the appearance of up to fifteen symbols. If more than fifteen symbols are needed, the first fifteen are repeated.

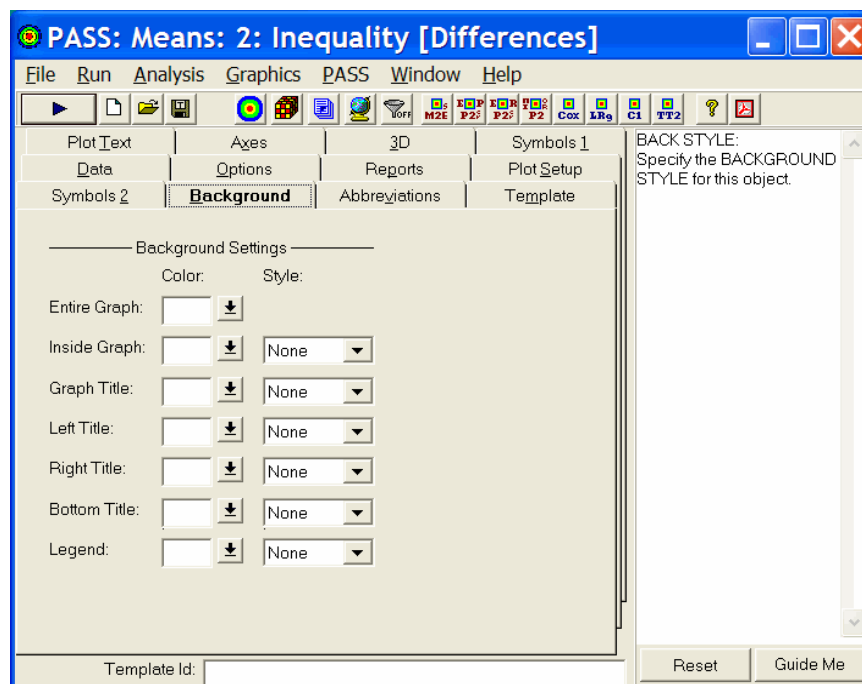


## Symbol

These options specify the color, size, and shape of the plotting symbols.

## Background tab

These options specify the style and color of various portions of the graph.



## Abbreviations tab

These tabs specify the abbreviations that are used for the parameters in the titles of the plots. It is usually necessary to keep these abbreviations as short as possible since the title can only contain 80 characters.

The screenshot shows the 'PASS: Means: 2: Inequality [Differences]' window with the 'Abbreviations' tab selected. The window has a menu bar (File, Run, Analysis, Graphics, PASS, Window, Help) and a toolbar with various icons. The main area is divided into two columns of dropdown menus for selecting abbreviations for different parameters. On the right, there is a text area for entering a custom abbreviation, and at the bottom, there are 'Reset' and 'Guide Me' buttons.

Parameter	Abbreviation
Test Title:	Alpha:
T Test	Alpha
Mean1:	Beta:
M1	Beta
Mean2:	Sample Size Group 1:
M2	N1
Standard Deviation 1:	Sample Size Group 2:
S1	N2
Standard Deviation 2:	Power:
S2	Power

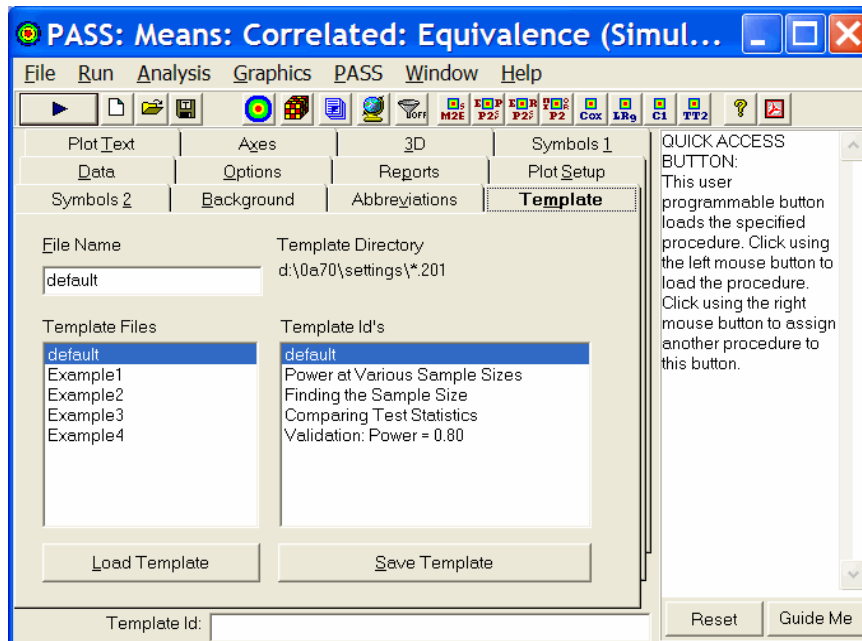
ABBREVIATION:  
Enter the character(s) to be used for this parameter on the plots in labels and titles. Remember that plot titles are limited to just 80 characters, so brevity is important here.

Template Id:

Reset Guide Me

## Template tab

This tab displays the options necessary to load and save templates. The *template* refers to the settings of all options under all tabs. Templates are stored in files in the *Settings* subdirectory.



### File Name

This box contains the name of the template file. Standard file naming conventions must be followed. A two or three character file extension is supplied by *PASS*. All you need to enter is the file name.

### Template Files

Select a template file from this list of previously saved templates. The file's name will be entered in the File Name box at the top of this window. You can then press the *Load Template* button to load the settings stored in this file into this template or the *Save Template* button to save the settings.

### Template Id's

This box contains a list of the template id's corresponding to the template name. Remember that the template id is entered in the text box at the bottom of the window.

### Load Template

This button will load the file whose name shows in the File Name box.

### Save Template

This button will save the settings of the all options to the template file named in the File Name box.

## CHAPTER 5

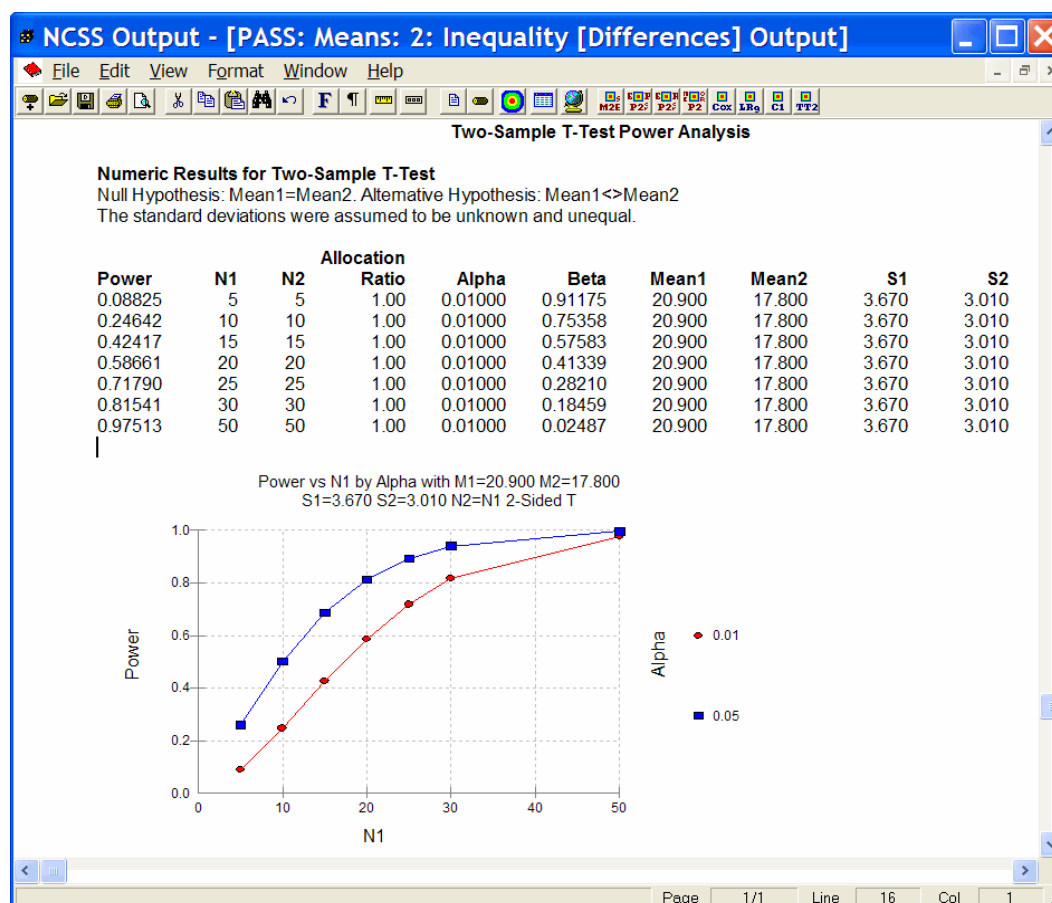
# Output Window

## Introduction

*PASS* sends all statistics and graphics output to its built-in word processor from where they can be viewed, edited, printed, or saved. Reports and graphs are saved in rich text format (RTF). Since RTF is a standard document transfer format, these files may be loaded directly into your word processor for further processing. You can also cut and paste data onto an **NCSS** datasheet for further analysis. This chapter covers the basics of our built-in word processor.

## Viewing the output

The output of the Example1 template of the Two-Sample T-Test program is shown below. The output window is in full-screen mode. The screen will look similar to this. Note that the actual size of your screen depends on the resolution of your monitor, so it may vary.



# Documents

The *PASS* word processor maintains two documents: *Output* and *Log*. Although both of these documents allow you to view your data, the *Output* document serves as a viewer while the *Log* document serves as a recorder.

You can load additional documents as well. For example, you might want to view the output from a previous analysis to compare the results with the current analysis. To do this, you open a third document that is actually the log file from a previous analysis.

All *PASS* documents are stored in the RTF format. This is a common format that is used by most word processors, including MS-Word and MS-Write. When you save a *PASS* report, you will be able to load that report directly into your own word processor. All text, formatting, and graphics will appear in your word processor ready for further editing. You can then save the document in your word processor's native format. In this way, you can easily transfer the output of a *PASS* procedure to almost any format you desire.

## Output Document

The Output document displays the output report from the current analysis. Whenever you run a *PASS* procedure, the resulting reports and graphs are displayed in the Output document. Each new run clears the existing Output document, so if you want to save a report, you must do so before running the next report.

The Output document provides four main functions: display, print, save to the Log document, and save as an RTF file.

## Log Document

The Log document provides a place to store a permanent record of your analysis. Since the Output document is erased by each new analysis, you need a place to store your permanent work. The Log document serves this purpose. When you have a report or graph that you want to keep, copy it from the Output document to the Log document.

The Log document provides four main word processing functions: load, display and edit, print, and save. When you load a file into the Log document, you can add new output to it. In this way, you can record your work on a project in a single file, even though your work on that project is spread out over several days.

# File Menu

The File Menu is used for opening, saving, and printing *PASS* word processor files. All options apply to the currently active document (the document whose title bar is selected). We will now discuss each of the options on this menu.

## New

This option opens an empty document. You might use this when you want to make notes about your analysis.

## New Log

This option opens an empty log document. You might use this when you want to start a new project.



## Open

This option opens an existing file. When this item is selected, the Open Report File dialog box appears. Note that no connection is maintained between a loaded file and its image on the disk. If you make changes to a file, you must save those changes to the disk.

## Open Log

This option opens an existing log file. When this item is selected, the Open Report File dialog box appears. The requested file is loaded into the Log document. Note that no connection is maintained between a loaded file and its image on the disk. If you make changes to a file, you must save those changes to the disk.

You might use this option when you want to continue using a certain file as the Log file.

## Toggle Auto-Log

When Auto-Log is on, the contents of the Output document are automatically copied to the Log document. The Output document remains unchanged. If you want to keep a copy of all the output that has been placed in the Log document, you will still need to save it manually.

This function allows you to automatically save all output for further use.

## Add Output to Log

Selecting this option copies the contents of the Output document to the Log document. The Output document remains unchanged. This allows you to save the current output document for further use.

## Save As

This option lets you save the contents of the currently active document to a designated file using the RTF format. Note that only the active document is saved. Also note that all file names should have the “RTF” extension so that other systems can recognize their format.

## Printer Setup

This option allows you to set printing options on your printer.

## Print Preview

This option displays the output report as it will appear on the printed page. Use it to preview your report before printing it out.

## Print...

This option lets you print the entire document or a range of pages. When you select this option, a Print Dialog box will appear that lets you control which pages are printed.

## Close Document

Minimizes the document that is currently being viewed. Note that this option does not clear the document, it simply minimizes it.

## Exit

This option exits the *NCSS/PASS* system. All documents and databases are closed.

# Edit Menu

This menu contains options that let you edit a document.

## Undo

This item reverses the last edit action. It is particularly useful for replacing something that was accidentally deleted.

## Cut

This item copies the currently selected text to the Windows clipboard and erases it from the document. You can paste the information from the clipboard to a different location in the current document, into another document, into a datasheet in the spreadsheet, or into another application. The selected text is erased.

## Copy

This item copies the currently selected text from the document to the Windows clipboard. You can paste this information from the clipboard to a different location in the current document, into another document, into a datasheet in the spreadsheet, or into another application. The selected text is not modified.

## Paste

This item copies the contents of the clipboard to the current document at the insertion point. This command is especially useful for moving selected information from the Output document to the Log document.

## Select All

This item selects the entire document. Although you can select a portion of the document using the mouse or a shift-arrow key, this is much faster if you want to select the entire document.

## Toggle Page Break

Changes the status of the page break on the line at which the insertion point resides. If a page break exists (shown by a horizontal line), it is removed. If a page break does not currently exist at that point, one is added.

Note that *PASS* does not repaginate your document for you. Once you make changes, it will be up to you to repaginate your document.

## Find

This item opens the Search dialog box. You can specify text that you want to search for. This is especially useful when you are looking for a certain topic or data value in a large report.

## Find Next

This item continues finding the text you entered in the Search Dialog box.

## Replace

This item opens the Search and Replace dialog box. This allows you to make repetitive changes. For example, you might want to change the name of one of the variables to a more useful name.

## Goto Section

This item does not modify the document. Instead, it lets you reposition the insertion point to one of the major topics. When *PASS* runs a procedure, it stores the major report topics in this list box. You can quickly position the view to a desired topic using this screen.

## View Menu

The *View Menu* lets you designate which editing tools you want to use.

### Ruler

This option controls whether the ruler and the tabs bar are displayed. The ruler displays the physical dimensions of the document. The tabs bar, found just below the ruler bar, lets you set the margins and tabs of your document. Only the currently selected part of your document is affected by a change in the tabs and margins.

### Format Toolbar

This option controls whether the Format Toolbar is displayed. The function of each of the buttons is discussed below.

### Status Bar

This option controls whether the Status Bar is displayed at the bottom of the output window.

### Show All

Selecting this menu item causes the Ruler, Tabs Bar, Format Toolbar, and Status Bar to be displayed.

### Hide All

Selecting this menu item causes the Ruler, Tabs Bar, Format Toolbar, and Status Bar to be hidden. This gives you more screen space to view your output.

### Redraw

Occasionally the Output Window becomes cluttered. If this happens, select this option to redisplay the output.

## Format Menu

This menu lets you set the format for a selected block of text.

### Font

This option displays the Replace Font dialog box, which lets you specify the font and style of the selected text.

### **Paragraph**

This option displays the Paragraph dialog box, which lets you specify the tabs and margins of the selected text.

### **Format Markers**

Indicates whether the (usually hidden) tab arrows and the end-of-paragraph marks are displayed in the document. Note that these characters are never printed.

## **Window Menu**

This menu lets you designate how you want the documents arranged on the screen and which window you want displayed on top of your output desktop.

### **Cascade**

This item arranges the documents in a cascading display from the upper left to the lower right of the screen.

### **Tile Horizontally**

This item arranges the documents horizontally down the word processor window.

### **Tile Vertically**

This item arranges the documents vertically across the word processor window.

### **Arrange Icons**

When a document is minimized, it is represented as an icon at the bottom of the word processor window. This option arranges all document icons. It is usually applied when the word processor window has been resized.

### **Current Output**

This item causes the Output window to be displayed.

### **Log**

This item causes the Log window to be displayed.

### **View Data (Spreadsheet)**

Causes the Spreadsheet window to be displayed.

### **View Navigator**

Causes the *NCSS/PASS* Navigator window to be displayed.

# Help Menu

This menu controls the display of the serial numbers and help system.

## Help

This item displays the help system.

## About

This item displays the release date and version number of your software.

## Serial Numbers

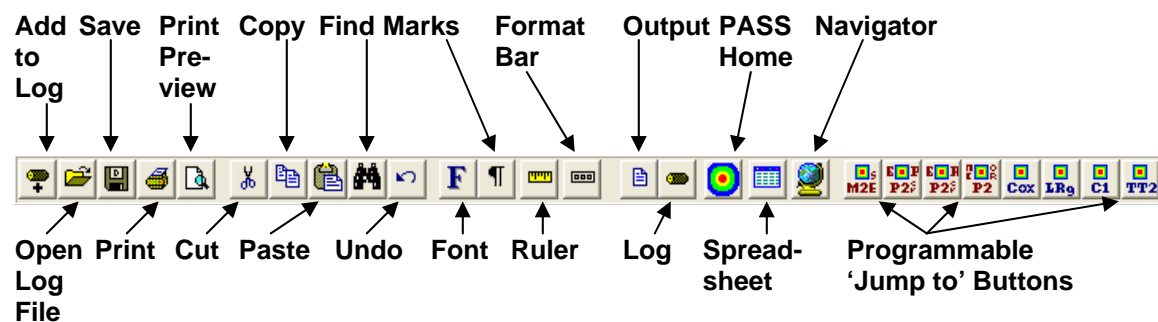
This item displays the *NCSS/PASS* Registration window where your serial numbers can be modified.

## View PDF File...

This option loads the PDF file that documents this procedure. This file may then be printed.

# Toolbar

The toolbar is a series of small buttons that appear just below the menus at the top of the procedure window. Each of these buttons provides quick access to a menu item.



## CHAPTER 6

# Introduction to Power Analysis

## Overview

A statistical test's *power* is the probability that it will result in statistical significance. Since statistical significance is the desired outcome of a study, planning to achieve high power is of prime importance to the researcher. Because of its complexity, however, an analysis of power is often omitted.

**PASS** calculates statistical power and determines sample sizes. It does so for a broad range of statistical techniques, including the study of means, variances, proportions, survival curves, correlations, bioequivalence, analysis of variance, log rank tests, multiple regression, and contingency tables.

**PASS** was developed to meet several goals, including ease of learning, ease of use, accuracy, completeness, interpretability, and appropriateness. It lets you study the influence of sample size, effect size, variability, significance level, and power on your statistical analysis.

# Brief Introduction to Power Analysis

Statistical power analysis must be discussed in the context of *statistical hypothesis testing*. Hence, this discussion starts with a brief introduction to statistical hypothesis testing, paying particular attention to topics that relate to power analysis and sample size determination. Although the theory behind hypothesis testing is general, its concepts can be reviewed by discussing simple case: testing whether a proportion is greater than a known standard.

Following the usual terminology of statistical hypothesis testing, define two complementary hypotheses

$$H_0: P \leq P_0 \text{ vs. } H_1: P > P_0$$

where  $P$  is the response proportion in the population of interest and  $P_0$  is the known standard value.

$H_0$  is called the *null hypothesis* because it specifies that the difference between the two proportions is zero (null).

$H_1$  is called the *alternative hypothesis*. This is the hypothesis of interest to us. Our motivation for conducting the study is to provide evidence that the alternative (or research) hypothesis is true. We do this by showing that the null hypothesis is unlikely—thus establishing that the alternative hypothesis (the only possibility left) is likely.

Outcomes from a statistical test may be categorized as follows:

1. Reject  $H_0$  when  $H_0$  is true. That is, conclude that  $H_0$  is unlikely when it is true. This constitutes a decision error known as the *Type-I error*. The probability of this error is alpha ( $\alpha$ ) and is often referred to as the *significance level* of the hypothesis test.
2. Do not reject  $H_0$  when  $H_0$  is false. That is, conclude that  $H_0$  is likely when it is false. This constitutes a decision error known as the *Type-II error*. The probability of this error is beta ( $\beta$ ). *Power* is  $1 - \beta$ . It is the probability of rejecting  $H_0$  when it is false.
3. Reject  $H_0$  when  $H_0$  is false. This is a correct decision.
4. Do not reject  $H_0$  when  $H_0$  is true. This is also a correct decision.

The basic steps in conducting a study that is analyzed with a hypothesis test are:

1. Specify the statistical hypotheses,  $H_0$  and  $H_1$ .
2. Run the experiment on a given number of subjects.
3. Calculate the value of a test statistic, such as the sample proportion.
4. Determine whether the sample values favor  $H_0$  or  $H_1$ .



## Binomial Probability Table

In the current example, suppose that a random sample of ten individuals is selected, that is  $N = 10$ . The number of individuals,  $R$ , with the characteristic of interest is counted. Hence,  $R$  is the test statistic. A table of binomial probabilities gives the probability that  $R$  takes on each of its eleven possible values for various values for  $P$ .

<b>R</b>	<b>P</b>								
	<b>0.1</b>	<b>0.2</b>	<b>0.3</b>	<b>0.4</b>	<b>0.5</b>	<b>0.6</b>	<b>0.7</b>	<b>0.8</b>	<b>0.9</b>
0	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000
1	0.387	0.376	0.121	0.040	0.010	0.002	0.000	0.000	0.000
2	0.194	0.302	0.233	0.121	0.044	0.011	0.001	0.000	0.000
3	0.057	0.201	0.267	0.215	0.117	0.042	0.009	0.001	0.000
4	0.011	0.088	0.200	0.251	0.205	0.111	0.037	0.006	0.000
5	0.001	0.026	0.103	0.201	0.246	0.201	0.103	0.026	0.001
6	0.000	0.006	0.037	0.111	0.205	0.251	0.200	0.088	0.011
7	0.000	0.001	0.009	0.042	0.117	0.215	0.267	0.201	0.057
8	0.000	0.000	0.001	0.011	0.044	0.121	0.233	0.302	0.194
9	0.000	0.000	0.000	0.002	0.010	0.040	0.121	0.376	0.387
10	0.000	0.000	0.000	0.000	0.001	0.006	0.028	0.107	0.349

Let us discuss in detail the interpretation of the values in this table for the simple case in which a coin is flipped ten times and the number of heads is recorded. The column parameter  $P$  is the probability of obtaining a head on any one toss of the coin. When dealing with coin tossing, one would usually set  $P = 0.5$ , but this does not have to be the case. The row parameter  $R$  is the number of heads obtained in ten tosses of a coin.

The body of the table gives the probability of obtaining a particular value of  $R$ . One way to interpret this probability value is as follows: conduct a simulation in which this experiment is repeated a million times for each value of  $P$ . Using the results of this simulation, calculate the proportion of experiments that result in each value of  $R$ . This proportion is recorded in this table. For example, when the probability of obtaining a head on a single toss of a coin is 0.5, ten flips of a coin would result in five heads 24.6% of the time. That is, as the procedure is repeated (flipping a coin ten times) over and over, 24.6% of the outcomes would be five heads.

## Calculating the Significance Level, Alpha

We will now explain how the above table is used to set the significance level (the probability of a type-I error) to a pre-specified value. Recall that a type-I error occurs when an experiment results in the rejection of the null hypothesis when, in fact, the null hypothesis is true. By studying the table, the impact of using different *rejection regions* can be determined. A rejection region is a simple rule that states which values of the test statistic will result in the null hypothesis being rejected.

For example, suppose we want to test  $P_0 = 0.5$ . That is, the null hypothesis is that  $P = 0.5$  and the alternative hypothesis is that  $P > 0.5$ . Suppose the rejection region is  $R$  equal to 8, 9, or 10. That is,  $H_0$  is rejected if  $R = 8, 9, \text{ or } 10$ . From the above table, the probability of obtaining 8, 9, or 10 heads in 10 tosses when  $P = 0.5$  is calculated as follows:

$$\begin{aligned} P(R = 8, 9, 10 \mid P = 0.5) &= P(R = 8 \mid P = 0.5) + P(R = 9 \mid P = 0.5) + P(R = 10 \mid P = 0.5) \\ &= 0.044 + 0.010 + 0.001 = 0.055. \end{aligned}$$

That is, 5.5% of these coin tossing experiments using this decision rule result in a type-I error. By setting the rejection criterion to  $R = 8, 9, \text{ or } 10$ , alpha has been set to 0.055.

It is extremely important to understand what alpha means, so we will go over its interpretation again. If the probability of obtaining a head on a single toss of a coin is 0.5, then 5.5% of the experiments that use the rejection criterion of  $R = 8, 9, \text{ or } 10$  will result in the false conclusion that  $P > 0.5$ .

The key features of this definition that are often overlooked by researchers are:

1. **The value of alpha is based on a particular value of  $P$ .** Note that we used the assumption “if the probability of obtaining a head is 0.5” in our calculation of alpha. Hence, if the actual value of  $P$  is 0.4, our calculations based on the assumption that  $P$  is 0.5 are wrong. Mathematicians call this a conditional probability since it is based on the condition that  $P$  is 0.5. Alpha is 0.055 if  $P$  is 0.5.

Often, researchers think that setting alpha to 0.05 means that the probability of rejecting the null hypothesis is 0.05. Can you see what is wrong with this statement? They have forgotten to mention the key fact that this statement is based on the assumption that  $P$  is 0.5!

2. **Alpha is a statement about a proportion in multiple experiments.** Alpha tells us what percentage of a large number of experiments will result in 8, 9, or 10 heads. Alpha is a statement about what to expect from future experiments. It is not a statement about  $P$ . Occasionally, researchers conclude that the alpha level is the probability that  $P = 0.5$ . This is not what is meant. Alpha is not a statement about  $P$ . It is a statement about future experiments, given a particular value of  $P$ .

## Interpreting $P$ Values

The term *alpha value* is often used interchangeably with the term *p value*. Although these two terms are closely related, there is an important distinction between them. A *p value* is the largest value of alpha that would result in the rejection of the null hypothesis for a particular set of data. Hence, while the value of alpha is set during the planning of an experiment, the *p value* is calculated from the data after experiment has been run.

## Calculating Power and Beta

We will now explain how to calculate the power. Recall that power is the probability of rejecting a false null hypothesis. A false  $H_0$  means that  $P$  is some value other than  $P_0$ . In order to compute power, we must know the actual value of  $P$ .

Returning to our coin tossing example, suppose the actual value of  $P$  is 0.7. What is the power and beta value of this testing procedure? The decision rule is to reject the null hypothesis when  $R$  is 8, 9, or 10. From the above probability table, the probability of obtaining 8, 9, or 10 heads in 10 tosses of a coin when probability of a head is actually 0.7 is

$$\begin{aligned} P(R = 8, 9, 10 \mid P = 0.7) &= P(R = 8 \mid P = 0.7) + P(R = 9 \mid P = 0.7) + P(R = 10 \mid P = 0.7) \\ &= 0.233 + 0.121 + 0.028 \\ &= 0.382 \end{aligned}$$

This is the power. The value of a type-II error is  $1.000 - 0.382$ , which is 0.618. That is, if  $P$  is 0.7, then 38.2% of these coin tossing experiments will reject  $H_0$ , while 61.8% of them will result in a type-II error.

It is extremely important to understand what beta means, so we will go over its interpretation again. If the probability of obtaining a head on the toss of a coin is 0.7, then 61.8% of the experiments that use the rejection criterion of  $R = 8, 9$ , or 10 will result in the false conclusion that  $P = 0.5$ .

The key features of this definition that are often overlooked by researchers are:

1. **The value of beta is based on a particular value of  $P$ .** Note that we used the assumption “if the probability of obtaining a head is 0.7” in our calculation of beta. Hence, if the actual value of  $P$  is 0.6, our calculation based on the assumption that  $P$  was 0.7 is wrong.
2. **Beta is a statement about the proportion of experiments.** Beta tells us what percentage of a large number of experiments will result in 8, 9, or 10 heads. Beta is a statement about what we can expect from future experiments. It is not a statement about  $P$ .
3. **Beta depends on the value of alpha.** Since the rejection region (8, 9, or 10 heads) depends on the value of alpha, beta depends on alpha.
4. **You cannot make both errors at the same time.** A type-II error can only occur when a type-I error did not occur, and vice versa.

## Specifying Alternative Values of the Parameters

We have noted a great deal of confusion about specifying the values of the parameters under the alternative hypothesis. The alternative hypothesis is usually that the value of one parameter is different from another. The hypothesis does not usually specify how different. It simply gives the direction of the difference. The power is calculated at specified alternative values. These values should be considered as values at which the power is calculated, not as the true value.

## Effect Size

The *effect size* is the size of the change in the parameter of interest that can be detected by an experiment. For example, in the coin tossing example, the parameter of interest is  $P$ , the probability of a head. In calculating the sample size, we would need to state what the baseline probability is (probably 0.5) and how large of a deviation from  $P$  that we want to detect with our experiment. We would expect that it would take a much larger sample size to detect a deviation of 0.01 than it would to detect a deviation of 0.40.

Selecting an appropriate effect size is difficult because it is subjective. The question that must be answered is: what size change in the parameter would be of interest? Note that, in power analysis, the effect size is not the actual difference. Instead, the effect size is the change in the parameter that is of interest or is to be detected. This is a fundamental concept that is often forgotten after the experiment is run.

After an experiment is run that leads to non-significance, researchers often ask, “What was the experiment’s power?” and “How large of a sample size would have been needed to detect significance?” To compute the power or sample size, they set the effect size equal to the amount that was seen in their experiment. This is incorrect. When performing a power analysis after an experiment has completed, the effect size is still the change in the parameter that would be of interest to other scientists. It is not the change that was actually observed!

Often, the effect size is stated as a percentage change rather than an absolute change. If this is the case, you must convert the percentage change to an absolute change. For example, suppose that you are designing an experiment to determine if tossing a particular coin has exactly a 50% chance of yielding a head. That is,  $P_0$  is 0.50. Suppose your gambling friends are interested in whether a certain coin has a 10% greater chance. That is, they are concerned with the case where  $P$  is 0.55 or greater. The effect size is  $|0.50 - 0.55|$  or 0.05.

## Types of Power Analyses

There are several types of power analyses. Often, power analysis is performed during the design phase of a study to determine the sample size. This type of study would determine the value of  $N$  for set values of  $\alpha$  and  $\beta$ . Another type of power analysis is a post hoc analysis, which is done after the study is concluded. A post hoc analysis studies such questions as:

1. What sample size would have been needed to detect a specific effect size?
2. What is the smallest effect size that could be detected with this sample size?
3. What was the power of the test procedure?

These and similar questions may be answered using power analysis. By considering these kinds of questions after a study is concluded, you can gain important insights into how to make your research more efficient and effective.

## Nuisance Parameters

Statistical hypotheses usually make statements about one or more parameters from a set of one or more probability distributions. Often, the hypotheses leave other parameters of the probability distribution unspecified. These unspecified parameters are called ‘nuisance’ parameters.

For example, a common clinical hypothesis is that the response proportions of two drugs are equal. The null hypothesis is that the difference between these two drugs is zero. The alternative is that the difference is non-zero. Note that the actual values of the two proportions are not stated in the hypothesis—just their difference. The actual values of the proportions will be needed to compute the power. That is, different powers will result for the case when  $P_1 = 0.05$  and  $P_2 = 0.25$  and for the case  $P_1 = 0.50$  and  $P_2 = 0.70$ . In this example, the proportion difference ( $D = P_1 - P_2$ ) is the parameter of interest. The baseline proportion,  $P_1$ , is a nuisance parameter.

Another example of a nuisance parameter occurs when using the t-test to test whether the mean is equal to a particular value. When computing the power or sample size for this test, the hypothesis specifies the value of the mean. However, the value of the standard deviation is also required. In this case, the standard deviation is a nuisance parameter.

When performing a power analysis, you should state all your assumptions, including the values of any nuisance parameters that were used. When you do not have any idea as to reasonable values for nuisance parameters, you should use a range of possible values so that you can analyze how sensitive the results are to the values of the nuisance parameters. Also, do not be tempted to use the nuisance parameter’s value from a previous (or pilot) study. Instead, a reasonable strategy is to compute a confidence interval and use the confidence limit that results in the largest sample size.

## Choice of Test Statistics

Many hypothesis tests can be tested with a variety of test statistics. For example, statisticians often have to decide between the t-test and the Wilcoxon test when testing means. Similarly, when testing whether two proportions are equal, they have to decide whether to use a z-test or an exact test. If they choose a z-test, they have to decide whether to apply a continuity correction.

In most cases, each test statistic will have a different power. Thus, it should be obvious that you must compute the power of the test statistic that will be used in the analysis. A sample size based on the t-test will not be accurate for a nonparametric test.

The next question is usually “Which test statistic should I use?” You might say “The one that requires the smallest sample size.” However, other issues besides power must be considered. For example, consideration must be given to whether the assumptions of the test statistic will be met by the data. If your data is binary, it is probably unreasonable to assume that they are continuous.

These are simple principles, but they are often overlooked.

# Types of Hypotheses

Hypothesis tests work this way. If the null hypothesis is rejected, the alternative hypothesis is concluded to be true. However, if null hypothesis is not rejected, no conclusion is reached--the null hypothesis is not concluded to be true. The only way that a conclusion is reached is if the null hypothesis is rejected.

Because of this, it is very important that the null and alternative hypotheses be constructed so that the conclusion of interest is associated with the alternative hypothesis. That way, if the null hypothesis is rejected, the study reaches the desired conclusion.

There are several types of hypotheses. These include inequality, equivalence, non-inferiority, and superiority hypotheses. In the statistical literature, these terms are used with completely different meanings, so it is important to define what is meant by each. We have tried to adopt names that are associated with the alternative hypothesis, since this is the hypothesis of interest.

It is important to note that even though two sets of hypotheses may be similar, they often will have different power and sample size requirements. For example, an equivalence test (see below) appears to be the simple reverse of a two-sided test of inequality, yet the equivalence test requires a much larger sample size to achieve the same power as the inequality test. Hence, you cannot select the sample size for an inequality test and then later decide to run an equivalence test.

Each of the sections to follow will give a brief definition along with an example based on the difference between two proportions.

## Inequality Hypothesis

The term 'inequality' represents the classical one-sided and two-sided hypotheses in which the alternative hypothesis is simply that the two values are unequal. These hypotheses are called tests of superiority by Julious (2004), emphasizing the one-sided versions.

### Two-Sided

When the null hypothesis is rejected, the conclusion is simply that the two parameters are unequal. No statement is made about how different. For example, 0.501 and 0.500 are unequal, as are 0.500 and 0.800. Obviously, even though the former are different, the difference is not large enough to be of practical importance in most situations.

$$H_0: p_1 - p_2 = 0 \text{ vs. } H_1: p_1 - p_2 \neq 0 \text{ or } H_1: p_1 - p_2 < 0 \text{ or } p_1 - p_2 > 0$$

### One-Sided

These tests offer a little more information than the two sided tests since the direction of the difference is given. Again, no indication is made about how much higher (or lower) the superior value is to the inferior.

$$H_0: p_1 - p_2 \leq 0 \text{ vs. } H_1: p_1 - p_2 > 0 \text{ or } H_0: p_1 - p_2 \geq 0 \text{ vs. } H_1: p_1 - p_2 < 0$$

## Non-Inferiority Hypothesis

These tests are a special case of the one-sided inequality tests. The term ‘non-inferiority’ is used to indicate that one treatment is not worse than another treatment. That is, one proportion is not less than another proportion by more than a trivial amount called the ‘margin of equivalence’.

For example, suppose that a new drug is being developed that is less expensive and has fewer side effects than the standard drug. Producers must show that its effectiveness is no worse than the drug it is to replace.

When testing two proportions in which a higher proportion is better, the non-inferiority of treatment 1 as compared to treatment 2 is expressed as

$$H_0: p_1 - p_2 \leq -\delta \text{ vs. } H_1: p_1 - p_2 > -\delta \text{ or } H_0: p_1 \leq p_2 - \delta \text{ vs. } H_1: p_1 > p_2 - \delta$$

where  $\delta > 0$  is called the margin of equivalence. Note that when  $H_0$  is rejected, the conclusion is that the first proportion is not less than the second proportion by more than  $\delta$ .

Perhaps an example will help introduce this type of test. Suppose that the current treatment for a disease works 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the first questions that must be answered is whether the new treatment is as good as the current treatment. In other words, do at least 70% of subjects respond to the new treatment?

Because of the many benefits of the new treatment, clinicians are willing to adopt the new treatment even if it is slightly less effective than the current treatment. They must determine, however, how much less effective the new treatment can be and still be adopted. Should it be adopted if 69% respond? 68%? 65%? 60%? There is a percentage below 70% at which the difference between the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, it is decided that if a response of at least 63% is achieved, the new treatment will be adopted. The difference between these two percentages is called the *margin of equivalence*. The margin of equivalence in this example is 7% (which is ten percent of the original 70%).

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is at least 0.63. The statistical hypothesis to be tested is

$$H_0: p_1 - p_2 \leq -0.07 \text{ versus } H_1: p_1 - p_2 > -0.07$$

Notice that when the null hypothesis is rejected, the conclusion is that the response rate is at least 0.63. Note that even though the response rate of the current treatment is 0.70, the hypothesis test is about a response rate of 0.63. Also, notice that a rejection of the null hypothesis results in the conclusion of interest.

## Superiority Hypothesis

These tests are a special case of the one-sided inequality tests. The term ‘superiority’ is used to indicate that one treatment is better than another by more than a trivial amount called the ‘margin of equivalence’. For example, suppose that a new drug is being developed that is thought to have superior performance to the existing drug. Producers must show that its effectiveness is better than the drug it is to replace.

When testing two proportions in which a higher proportion is better, the superiority of treatment 1 over treatment 2 is expressed as

$$H_0: p_1 - p_2 \leq \delta \text{ vs. } H_1: p_1 - p_2 > \delta \text{ or } H_0: p_1 \leq p_2 + \delta \text{ vs. } H_1: p_1 > p_2 + \delta$$

where  $\delta > 0$  is called the margin of equivalence. Note that when  $H_0$  is rejected, the conclusion is that the first proportion is higher than the second proportion by more than  $\delta$ .

## Equivalence Hypothesis

The term ‘equivalence’ is used here to represent tests designed to show that response rates of two treatments do not differ by more than a trivial amount called the ‘margin of equivalence’. These tests are the reverse of the two-sided inequality test.

The typical set of hypotheses are

$$H_0: p_1 - p_2 \leq \delta_L \text{ or } p_1 - p_2 \geq \delta_U \text{ vs. } H_1: \delta_L \leq p_1 - p_2 \leq \delta_U$$

where  $\delta_L < 0$  and  $\delta_U > 0$  are called the *equivalence limits*.

Suppose 70% of subjects with a certain disease respond to a certain drug. The company that produces the drug has decided to open a new facility in another city. They must show that the drug produced in the new facility is equivalent (all most the same) as that produced in existing facilities. After thoughtful discussion with several clinicians and regulatory agencies, it is decided that if the response rate of the drug produced at the new facility is between 65% and 75%, the new facility will go into production. In this case, the *margin of equivalence* is 5%.

The statistical hypothesis to be tested is

$$H_0: |p_1 - p_2| \geq 0.05 \text{ vs. } H_1: |p_1 - p_2| < 0.05$$



## Chapter 7

# Proportions

## Introduction

This chapter introduces power analysis and sample size calculation for proportions. When the response is binary, the results for each group may be summarized as proportions. Usually, hypothesis tests are conducted to compare two proportions.

There are many issues that must be considered when designing experiments that use proportions. For example, will the proportions be analyzed directly, or will they be converted into differences, ratios, or odds ratios? Which test statistic will be used? Will the design use independent groups or will subjects be measured twice? Will the study have an active control?

The various answers to these and other questions result in a large number of situations. This chapter will introduce you to the issues that are common to all the tests of proportions.

## Designs

There are several experimental designs for comparing two proportions. You can use a one-sample design to compare a sample proportion to a specific value. You can compare proportions from two independent samples using independent, stratified, cluster-randomized, or group-sequential designs. You can compare two correlated proportions. And finally, you can compare several categories in a contingency table.

## Comparing Proportions

Two proportions may be compared by considering their difference, ratio, or odds ratio. Each of these cases may lead to different test statistics with different powers and sample size requirements.

Assume that  $p_1$  is the response proportion of the experimental group and  $p_2$  is the response proportion of the control (standard or reference) group. Mathematically, these alternative parameterizations are

<u>Parameter</u>	<u>Computation</u>
Difference	$\delta = p_1 - p_2$
Ratio	$\phi = p_1 / p_2$
Odds Ratio	$\psi = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$

Once you know  $p_1$  and  $p_2$ , you can calculate any of these values, and you can easily change from one parameterization to another. Thus, the choice of which parameter you use may seem arbitrary. However, since different parameterizations lead to different test statistics, the choice can lead to a different power and sample size. These parameterizations will be discussed next.

## Difference

The difference  $\delta = p_1 - p_2$  is perhaps the most common method of comparing two proportions. This parameter is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its use.

One difficulty occurs when the event of interest is rare. If a difference of 0.001 is reported for an event with a baseline probability of 0.40, we would dismiss this as being trivial. That is, there is usually little interest in a treatment that decreases the probability from 0.400 to 0.399. However, if the baseline probability of a disease is 0.002, a 0.001 decrease in the disease probability would represent a 50% reduction. Thus, the interpretation of the difference depends on the baseline probability of the event.

When planning studies, the value of  $p_2$  is usually known and the power is calculated at various values of  $\delta$ . The value of  $p_1$  is then calculated using  $p_1 = p_2 + \delta$ . Because of the requirement that  $0 < p_1 < 1$ , the values of  $\delta$  are restricted to the interval  $-p_2 < \delta < 1 - p_2$ , not  $-1 < \delta < 1$  as you might expect. Likewise, the values of  $p_2$  are restricted to  $0 < p_2 < 1 - \delta$  if  $\delta > 0$  and  $-\delta < p_2 < 1$  if  $\delta < 0$ .

Because typical values of  $\delta$  are usually between -0.2 and 0.2, the allowable values of  $p_2$  are restricted to be between 0.2 and 0.8. When the values of  $p_2$  are outside this range, it may be more convenient to use the odds ratio.

## Ratio

The (risk) ratio,  $\phi = p_1 / p_2$ , gives the relative change in the probability of the outcome under each of the hypothesized values. This parameter is direct and easy to interpret. To compare the ratio with the difference, examine the case where  $p_1 = 0.1437$  and  $p_2 = 0.0793$ . One should consider which number is more enlightening,  $\delta = -0.0644$ , or  $\phi = 55.18\%$ . In many cases, the relative change (the ratio) is easier to interpret than the absolute change (the difference).

When planning studies, the value of  $p_2$  is usually known and the power is calculated at various values of  $\phi$ . The value of  $p_1$  is then calculated using  $p_1 = p_2 \times \phi$ . Because of the requirement that  $0 < p_1 < 1$ , the values of  $\phi$  are restricted to the interval  $0 < \phi < 1 / p_2$ , not  $0 < \phi < \infty$  as you might expect. Likewise, the values of  $p_2$  are restricted to  $0 < p_2 < 1 / \phi$  if  $\phi > 1$  and  $0 < p_2 < 1$  if  $\phi < 1$ .

Because typical values of  $\phi$  are usually between 0.5 and 1.5, the values of  $p_2$  are restricted to be between 0 and 0.667. When the values of  $p_2$  are outside this range, it may be more convenient to use the odds ratio.

## Odds Ratio

Chances are usually communicated as long-term proportions or probabilities. In betting, chances are often given as odds. For example, the odds of a horse winning a race might be set at 10-to-1 or 3-to-2. Odds can easily be translated into probabilities, and vice versa. An odds of 3-to-2 means that the event is expected to occur three out of five times. That is, an odds of 3-to-2 (1.5) translates to a probability of winning of 0.60.

The odds of an event are calculated by dividing the event risk by the non-event risk. Thus the odds are

$$Odds_1 = \frac{p_1}{1 - p_1} \text{ and } Odds_2 = \frac{p_2}{1 - p_2}$$

For example, if  $p$  is 0.6, the odds are  $0.6/0.4 = 1.5$ . Rather than represent the odds as a decimal amount, it is re-scaled into whole numbers. Thus, instead of presenting the odds as 1.5-to-1, they present as 3-to-2.

Two odds could be compared by considering their difference, but it is more convenient in many situations to form their ratio. Thus, another way to compare proportions is to compute the ratio of their odds. The odds ratio is

$$\begin{aligned} \psi &= \frac{Odds_1}{Odds_2} \\ &= \frac{\frac{p_1}{1 - p_1}}{\frac{p_2}{1 - p_2}} \end{aligned}$$

Unlike the difference and the ratio, the odds ratio is not restricted by the value of  $p_2$ . The range of possible values of the odds ratio is  $-\infty < \psi < \infty$ . Because of the freedom in specifying the parameters, the odds ratio is a popular parameterization, even though it is not as easy to interpret as the difference and the ratio.

## Specifying the Proportions – Very Important!

It is important to understand the interpretation of  $p_1$  and  $p_2$  within *PASS*. Suppose  $p_1$  represents the proportion in the treatment group and  $p_2$  represents the proportion in the control group. In most hypothesis tests, these values are equal under the null hypothesis and different under the alternative hypothesis. Thus, under the null hypothesis, all that is needed is the value of  $p_1$  or  $p_2$ , but not both. Under the alternative hypothesis, both values are necessary. So, when the input screen asks for  $p_2$  and the difference, these values should be interpreted as follows. The value of  $p_2$  is actually the value of both  $p_1$  and  $p_2$  under  $H_0$ . Under  $H_1$ , the value of  $p_1$  is calculated from  $p_2$  and  $\delta$  using the formula  $p_1 = p_2 + \delta$ .

Also, it is important to understand what we mean by ‘under  $H_1$ ’ in the above discussion. Notice that  $H_1$  does not specify the exact value of  $p_1$ . Instead, it specifies a range of values. For example,  $H_1$  might be  $p_1 > p_2$  or  $p_1 - p_2 > \delta$ . However, even though  $H_1$  gives a range of values for  $p_1$ , the power is computed at a specific value of  $p_1$ .

Selecting an appropriate value for  $p_1$  must be done very carefully. We recommend the following approach. Select a value of  $p_1$  that represents the change from  $p_2$  that you want the experiment to detect. When you calculate a sample size, it is interpreted as the sample size necessary to detect a difference of at least  $p_1 - p_2$  when the significance level is  $\alpha$  and the power is  $1 - \beta$ .

The important point is that  $p_1$  is not the value you anticipate obtaining from an experiment. Instead, it is that value of  $p_1$  at which you want to compute the power. This is a very important distinction! This is why, when investigating the power after an experiment is run, we recommend that you do not simply plug in the values of  $p_1$  and  $p_2$  from that experiment. Rather, you use values that represent the size of the difference that you want to detect.

## Chapter 8

# Means

## Introduction

This chapter introduces power analysis and sample size calculation for tests that compare means. In many situations, the results for each treatment group are summarized as means. There are many issues that must be considered when designing experiments for comparing means. For example, are the means independent or correlated? Which test statistic to use? Will a parametric or nonparametric test be used? Are the data normally distributed? Are there more than two treatment groups? The answers to these and other questions result in a large number of situations.

## Specifying the Means

Assume that  $\mu_1$  is the mean of an experimental group and  $\mu_2$  is the mean of a control (standard or reference) group. Suppose  $\delta$  represents their difference. That is,  $\delta = \mu_1 - \mu_2$ . In most hypothesis tests, the null hypothesis ( $H_0$ ) is  $\delta = 0$  and the alternative hypothesis ( $H_1$ ) is  $\delta \neq 0$ . Since  $H_0$  assumes that  $\delta = 0$ , all that is really needed to compute the power is the value of  $\delta$  under  $H_1$ . So, when the input screen asks for  $\mu_1$  and  $\mu_2$ , these values should be interpreted as follows. The value of  $\mu_1$  is actually the value of both  $\mu_1$  and  $\mu_2$  under  $H_0$ . Under  $H_1$ , the values of  $\mu_1$  and  $\mu_2$  provide the value of  $\delta$  at which the power is calculated.

The above discussion is summarized in the following chart.

Input Parameter	Under $H_0$	Under $H_1$
Mean1	$\mu_1, \mu_2$	$\mu_1$
Mean2	ignored	$\mu_2$

Also, it is important to understand what we mean by ‘under  $H_1$ ’ in the above discussion.  $H_1$  defines a range of values for  $\delta$  at which the power can be computed. To compute the power, the specific values of  $\delta$  must be determined. Thus, there is not a single power value. Instead, there are an infinite number of power values possible, depending on the value of  $\delta$ .

Selecting an appropriate value for  $\mu_1$  must be done very carefully. We recommend the following approach. Select a value of  $\mu_1$  that represents the change from  $\mu_2$  that you want the experiment to detect. When you calculate a sample size, it is interpreted as the sample size necessary to detect a difference of at least  $\delta$  when the significance level is  $\alpha$  and the power is  $1 - \beta$ .

It is important to realize that  $\delta$  is not the value you anticipate obtaining from the experiment. Instead, it is that value of  $\delta$  at which you want to compute the power. This is a very important distinction! This is why, when investigating the power after an experiment is run, we recommend that you do not simply plug in the values of  $\mu_1$  and  $\mu_2$  from that experiment. Rather, you use values that represent the size of the difference that you want to detect.

## Specifying the Standard Deviation

Usually, statistical hypotheses about the means make no direct statement about the standard deviation. However, the standard deviation is a parameter in the normal distribution, so its value must be specified. For this reason, it is called a *nuisance parameter*.

Even though it is not of primary interest, an estimate of the standard deviation is necessary to perform a power analysis. Finding such an estimate is difficult not only because it is required before the data are available, but also because the physical interpretation of the standard deviation is vague. How do you estimate a quantity without data and without a clear understanding of what it is? This section will try to help.

## Understanding the Standard Deviation

The standard deviation has two general interpretations. First, it is similar to the average absolute difference between each observation and the mean. Second, it is the average absolute difference between every pair of observations.

The standard deviation of a population of values is calculated using the formula

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N}}$$

where  $N$  is the number of items in the population,  $X$  is the variable being measured, and  $\mu_X$  is the mean of  $X$ . This formula indicates that the standard deviation is the square root of an average of the squared differences between each value and the mean. The differences are squared to remove the sign so that negative values will not cancel out positive values. After summing up these squared differences and dividing by  $N$ , the square root is taken to put the result back in the original scale. Bottom line—the standard deviation can be thought of as the average absolute difference between the data values and their mean.

## Estimating the Standard Deviation

Our task is to find a rough estimate of the standard deviation to use in a power analysis. Several possible methods could be used. These include using the results of a previous study or a pilot study, using the range, using the coefficient of variation, etc. **PASS** includes a Standard Deviation Estimator procedure that will help you obtain a standard deviation estimate based on these methods. It is loaded from the PASS-Other menu. Remember that as the standard deviation increases, the power decreases. Hence, an increase in the standard deviation will cause an increase in the sample size. To be conservative in sample size calculation, you should use a large value for the standard deviation.

# Simulation

Most of the formulas used in *PASS* were derived by analytic methods. That is, based on a series of assumptions, a formula for the power and sample size is derived mathematically. This formula is then programmed and made available in *PASS*. Unfortunately, the formula is only as realistic as the assumptions upon which it is based. If the assumptions are inaccurate in a certain situation, the power calculations may also be inaccurate. An alternative to using analytic methods is to use *simulation* (or *Monte Carlo*) techniques. Because of the speed of modern computers, simulations can now be run in minutes that would have taken days or weeks only a few years ago.

In power analysis, *simulation* refers to the process of generating several thousand random samples that follow a particular distribution, calculating the test statistic from each sample, and tabulating the distribution of these test statistics so that the significance level and power of the procedure may be estimated.

The steps to a simulation study are

1. Specify how the study is carried out. This includes specifying the randomization procedure, the test statistic that is used, and the significance level that will be used.
2. 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.
3. 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.
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 significance level is the proportion of simulated samples in step 2 that lead to rejection. The power is the proportion of simulated samples in step 3 that lead to rejection.

## How Large Should the Simulation Be?

As the number of simulations is increased, the precision and running time of the simulation will be increased also. This section provides a method for estimating of the number simulations needed to achieve a given precision.

Each simulation iteration (or simulation) generates a binary outcome: either the null hypothesis is rejected or not. Thus, the significance level and power estimates each follow the binomial distribution. This knowledge makes it a simple matter to compute confidence intervals for the significance level and power values.

The following table gives one-half the width of a 95% confidence interval for the power when the estimated value is either 0.50 or 0.95.

Simulation Size M	Half-Width when Power = 0.50	Half-Width 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, as the simulation size is increased beyond 5000, there is only a small amount of additional accuracy achieved. Since most sample-size studies require an accuracy of within one or two percentage points, simulation sizes from 2000 to 10000 should be ample.

## You are Running Two Simulations

It is important to realize that when you run a simulation in *PASS*, you are actually running two separate simulations: one to estimate the significance level and the other to estimate the power. The significance-level simulation is defined by the input parameters labeled 'H0'. The power simulation is defined by the input parameters labeled 'H1'. Even though you have complete flexibility as to what distributions you use in each of these simulations, it usually makes sense to use the same distributions for both simulations—only changing the values of the means.

## Unequal Standard Deviations

One of the subtle problems that can make the results of a simulation study misleading is to specify unequal standard deviations unknowingly when you are investigating another feature, such as the amount of skewness. It is well known that if the standard deviations differ (a situation called heteroskedasticity), the accuracy of the significance level and power is doubtful. When investigating the power of the t or F tests in non-normal situations, care must be taken to insure that the standard deviations of the groups remain about the same. Otherwise, the effects of skewness and heteroskedasticity cannot be separated.



## Finding the Hypothesized Means

It is important to set the mean difference of each simulation carefully. In the case of analytic formulas, the mean difference is specified easily and directly. Usually, the mean difference is set to zero under the null hypothesis and to a non-zero value under the alternative hypothesis. You must make certain that you follow this pattern when setting up a simulation.

For most distributions, the means are set explicitly (the exception is the multinomial distribution, where this is impossible). Hence, for both the null and alternative simulations, it is relatively simple to calculate the mean difference. You must do this! We will now present two examples showing how this is done.

For the first example, consider the case of a simulation being run to compare two independent group means using the two-sample t-test. Suppose the *PASS* setup is as follows. Note that N(40 2) stands for a normal distribution with a mean of 40 and a standard deviation of 2.

<u>Distribution</u>	<u>PASS Input</u>	<u>Mean Value of Simulated Data</u>
Group 1 Distribution   H0	N(40 2)	40.0
Group 2 Distribution   H0	N(40 2)	40.0
Group 1 Distribution   H1	N(42 2)	42.0
Group 2 Distribution   H1	N(40 2)	40.0

The mean difference under H0 is  $40 - 40 = 0$ , which is as it should be. The mean difference under H1 is  $42 - 40 = 2$ . Hence, the power is being estimated for a mean difference of 2.

Next we will consider a more complicated example. Suppose the *PASS* setup is as follows. Note that N(40 2)[70];K(0)[30] specifies a mixture distribution made up of 70% from a normal distribution with a mean of 40 and a standard deviation of 2 and 30% from a constant distribution with a value of 30.

<u>Distribution</u>	<u>PASS Input</u>	<u>Mean Value of Simulated Data</u>
Group 1 Distribution   H0	N(40 2) [70];K(0)[30]	$40(0.7) + 30(0.3) = 37.0$
Group 2 Distribution   H0	N(40 2) [70];K(0)[30]	$40(0.7) + 30(0.3) = 37.0$
Group 1 Distribution   H1	N(42 2) [70];K(0)[30]	$42(0.7) + 30(0.3) = 38.4$
Group 2 Distribution   H1	N(40 2) [70];K(0)[30]	$40(0.7) + 30(0.3) = 37.0$

The mean difference under H0 is  $37.0 - 37.0 = 0$ , which is as it should be for the null hypothesis. The mean difference under H1 is  $38.4 - 37.0 = 1.4$ . Hence, the power is being estimated by simulation for a mean difference of 1.4.

You must always be aware of what the mean differences are under both the null and alternative hypotheses.

## Adjusting the Significance Level

When faced with the task of designing an experiment that will have a specific significance level for a situation that does not meet the usual assumptions, there are several possibilities.

1. A statistician could be hired to find an appropriate testing procedure.
2. A nonparametric test could be run that (hopefully) corrects for the implausible assumptions.
3. The regular parametric test could be run, relying on the test's 'robustness' to correct for the implausible assumptions.
4. A simulation study could be conducted to determine an appropriate adjustment to the significance level so that the actual significance level is at the required value.

We will now present an example of how to do the simulation adjustment alluded to in item 4, above.

The two-sample t-test is known to be robust to the violation of some assumptions, but it is susceptible to inaccuracies when the data contain outliers. A mixture of two normal distributions will be used to generate data with 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 zero and a standard deviation of 10. A simulation study using 10,000 iterations and a sample size of 100 per group produced the following results when the nominal significance level was set to 0.05.

<b>Nominal</b>	<b>Actual</b>	<b>Lower 95%</b>	<b>Upper 95%</b>	
<b><u>Alpha</u></b>	<b><u>Alpha</u></b>	<b><u>Confidence</u></b>	<b><u>Confidence</u></b>	<b><u>Power</u></b>
0.050	0.045	0.041	0.049	0.816
0.055	0.051	0.047	0.055	0.843
0.060	0.057	0.053	0.062	0.835

The actual alpha level of the t-test is 0.045, which is below the target value of 0.05. When the nominal alpha level is increased to 0.055, the actual alpha is 0.051—close to the desired level of 0.05. Hence, an adjustment could be applied as follows. Analyze the data with the two-sample t-test even though they contain outliers. However, instead of using an alpha of 0.050, use an alpha of 0.055. When this is done, the simulation shows that the actual alpha will be at the desired 0.05 level.

There is one limitation to this method: the resulting test procedure is not necessarily efficient. That is, it may be possible to derive a testing procedure that is more efficient (requires a smaller sample size to achieve the same power). For example, in this example, a test based on the trimmed mean may be more efficient in the presence of outliers. However, if you do not have the time or ability to derive an alternative test, this adjustment allows you to obtain reasonable testing procedure that achieves a desired significance level and power.

## Chapter 100

# One Proportion – Inequality

## Introduction

The *One-Sample Proportion Test* is used to assess whether a population proportion is significantly different from a hypothesized value. This is called the hypothesis of *inequality*. The hypotheses may be stated in terms of the proportions, their difference, their ratio, or their odds ratio.

For example, suppose that the current treatment for a disease cures 62% of all cases. A new treatment method has been proposed and studied. In a sample of 80 subjects with the disease that were treated with the new method, 63 were cured. Do the results of this study support the claim that the new method has a higher response rate than the existing method?

This procedure calculates sample size and statistical power for testing a single proportion using either exact or approximate tests. Results are based on exact calculations using the binomial and hypergeometric distributions. Because the analysis of several different test statistics is available, their statistical power may be compared to find the most appropriate test for a given situation.

Some sample size programs use the normal approximation to the binomial distribution for power and sample size estimates. This approximation is useful for rough hand calculations, but more accurate results are easily obtainable with today's software. When the normal approximation to the binomial is used, issues such as the need for continuity correction come into play. We avoid these issues by calculating exact results. Programs that use these approximations will often give different answers. Our calculations are exact, not approximate.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Binomial Model

A binomial variable should exhibit the following four properties:

1. The variable is binary--it can take on one of two possible values.
2. The variable is observed a known number of times. Each observation or replication is called a Bernoulli trial. The number of replications is  $n$ . The number of times that the outcome of interest is observed is  $r$ . Thus  $r$  takes on the possible values 0, 1, 2, ...,  $n$ .
3. The probability,  $P$ , that the outcome of interest occurs is constant for each trial.
4. The trials are independent. The outcome of one trial does not influence the outcome of the any other trial.

A binomial probability is calculated using the formula

$$b(r; n, P) = \binom{n}{r} P^r (1 - P)^{n-r}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

## The Hypergeometric Model

When samples are taken without replacement from a population of known size,  $N$ , the *hypergeometric* distribution should be used in place of the binomial distribution. The properties of a variable that is distributed according to the hypergeometric distribution are

1. The variable is binary--it can take on one of two possible values.
2. The variable is observed a known number of times. Each observation or replication is called a Bernoulli trial. The number of replications is  $n$ . The number of times that the outcome of interest is observed is  $r$ . Thus  $r$  takes on the possible values 0, 1, 2, ...,  $n$ .
3. The total number of items is  $N$ . The proportion of items with the characteristic of interest is  $P$ .

The hypergeometric probability of obtaining exactly  $r$  of  $n$  items with the characteristic of interest is calculated using

$$h(r; N, n, P) = \frac{\binom{NP}{r} \binom{N - NP}{n - r}}{\binom{N}{n}}$$

Note that the quantity  $NP$  is rounded to the nearest integer.

# Test Procedure

The testing procedure is as follows. Let  $P$  represent the true probability that an item selected at random from a population will have a characteristic of interest.

## 1. State the hypotheses.

$H_0$  is the null hypothesis that the proportion is  $P_0$ .

$H_1$  is the alternative hypothesis that the proportion is  $P_1$ .

Three sets of statistical hypotheses may be formulated:

1.  $H_0: P = P_0$  versus  $H_1: P = P_1 \neq P_0$ . This hypothesis results in a two-tailed test.
2.  $H_0: P \leq P_0$  versus  $H_1: P = P_1 > P_0$ . This hypothesis results in a one-tailed test.
3.  $H_0: P \geq P_0$  versus  $H_1: P = P_1 < P_0$ . This hypothesis results in a one-tailed test.

## 2. Find the critical value.

For an upper-tailed test with a given sample size find the critical value,  $P_c$ , based on the binomial (or hypergeometric) distribution, so that the probability of rejecting  $H_0$  when  $H_0$  is true is equal to a specified significance level,  $\alpha$ .

## 3. Evaluate the sample.

Select a sample of  $n$  items from the population and compute the sample proportion,  $p = r / n$ . If  $p > P_c$  then reject the null hypothesis that  $P = P_0$  in favor of an alternative hypothesis that  $P = P_1 > P_0$ .

## 4. Calculate the power.

The power is the probability of rejecting  $H_0$  when the true proportion is  $P_1$ . That is, the power is the probability that  $p > P_c$  calculated from a binomial (or hypergeometric) distribution in which  $P = P_1$ .

Similar steps are used for the lower-tail and two-tailed tests.

## Test Statistics

Many different test statistics have been proposed for testing a single proportion. Most of these were proposed before computers or hand calculators were widely available. Although these legacy methods are still presented in textbooks, their power and accuracy should be compared against modern exact methods before they are adopted for serious research. To make this comparison easy, the power and significance of several tests of a single proportion are available in this procedure.

### Exact Binomial Test

The test statistic is  $r$ , the number of successes in  $n$  trials. This test should be the standard against which other test statistics are judged. The significance level and power are computed by enumerating the possible values of  $r$ , computing the probability of each value, and then computing the corresponding value of the test statistic. Hence the values that are reported in the output for these tests are exact, not approximate.

### Z Tests

Several  $z$  statistics have been proposed that use the central limit theorem. This theorem states that for large sample sizes, the distribution of the  $z$  statistic is approximately normal. All of these tests take the following form:

$$z = \frac{p - P0}{s}$$

Although these  $z$  tests were developed because the distribution of  $z$  is approximately normal in large samples, the actual significance level and power can be computed exactly using the binomial distribution.

We include four  $z$  tests which are based on two methods for computing  $s$  and whether a continuity correction is applied.

#### Z Test using s(P0)

This test statistic uses the value of  $P0$  to compute  $s$ .

$$z_1 = \frac{p - P0}{\sqrt{P0(1 - P0) / n}}$$

**Z Test using s(P0) with Continuity Correction**

This test statistic is similar to the one above except that a continuity correction is applied to make the normal distribution more closely approximate the binomial distribution.

$$z_2 = \frac{(p - P0) + c}{\sqrt{P0(1 - P0) / n}}$$

where

$$c = \begin{cases} -\frac{1}{2n} & \text{if } p > P0 \\ \frac{1}{2n} & \text{if } p < P0 \\ 0 & \text{if } |p - P0| < \frac{1}{2n} \end{cases}$$

**Z Test using s(P-hat)**

This test statistic uses the value of  $p$  to compute  $s$ .

$$z_3 = \frac{p - P0}{\sqrt{p(1 - p) / n}}$$

**Z Test using s(P-hat) with Continuity Correction**

This test statistic is similar to the one above except that a continuity correction is applied to make the normal distribution more closely approximate the binomial distribution.

$$z_4 = \frac{(p - P0) + c}{\sqrt{p(1 - p) / n}}$$

where

$$c = \begin{cases} -\frac{1}{2n} & \text{if } p > P0 \\ \frac{1}{2n} & \text{if } p < P0 \\ 0 & \text{if } |p - P0| < \frac{1}{2n} \end{cases}$$

**T Test**

The one-sample t-test may be applied to this design. This is accomplished by considering the  $n$  trials as the outcomes of a numeric variable in which a success is coded as a '1' and a failure is coded as a '0'. The standard t-test may then be computed on these data values.

## Parameterizations of the Proportions

There are several ways to specify the proportions under the null and the alternative hypotheses. The most direct is to simply give values for  $P_0$  and  $P_1$ . However, it is often more meaningful to specify  $P_0$  and then specify the alternative as the difference, the ratio, or the odds ratio. The value of  $P_1$  is calculated from these values.

Mathematically, these alternative parameterizations are

<u>Parameter</u>	<u>Computation</u>
Difference	$\delta = P_1 - P_0$
Ratio	$\phi = P_1 / P_0$
Odds Ratio	$\psi = \frac{P_1 / Q_1}{P_0 / Q_0} = \frac{P_1 Q_0}{P_0 Q_1}$

### Difference

The (risk) difference,  $\delta = P_1 - P_0$ , is perhaps the most direct method of comparison between the two proportions. This parameter is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One interpretation difficulty occurs when the event of interest is rare. If a difference of 0.001 is reported for an event with a baseline probability of 0.40, we would dismiss this as being trivial. That is, there is usually little interest in a treatment that decreases the probability from 0.400 to 0.399. However, if the baseline probability of a disease is 0.002, a 0.001 decrease in the disease probability would represent a reduction of 50%. The interpretation depends on the baseline probability of the event.

### Ratio

The (risk) ratio,  $\phi = P_1 / P_0$ , gives the relative change in the probability of the outcome under each of the hypothesized values. This parameter is direct and easy to interpret. To compare the ratio with the difference, examine the case where  $P_0 = 0.1437$  and  $P_1 = 0.0793$ . One should consider which number is more enlightening, the difference of -0.0644, or the ratio of 55.18%. In many cases, the ratio communicates the change in proportion in a manner that is more appropriate than the difference.



## Odds Ratio

Chances are usually communicated as long-term proportions or probabilities. In betting, chances are often given as odds. For example, the odds of a horse winning a race might be set at 10-to-1 or 3-to-2. Odds can easily be translated into probability. An odds of 3-to-2 means that the event is expected to occur three out of five times. That is, an odds of 3-to-2 (1.5) translates to a probability of winning of 0.60.

The odds of an event are calculated by dividing the event risk by the non-event risk. Thus the odds are

$$O1 = \frac{P1}{1 - P1} \text{ and } O0 = \frac{P0}{1 - P0}$$

For example, if P1 is 0.60, the odds are  $0.60/0.4 = 1.5$ . Rather than represent the odds as a decimal amount, it is re-scaled into whole numbers. Thus, instead of saying the odds are 1.5-to-1, we say they are 3-to-2.

Thus, another way to compare proportions is to compute the ratio of their odds. The odds ratio of two proportions is

$$\begin{aligned} \psi &= \frac{O1}{O0} \\ &= \frac{\frac{P1}{1 - P1}}{\frac{P0}{1 - P0}} \end{aligned}$$

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are  $P0$ ,  $P1$ ,  $\text{Alpha}$ ,  $\text{Beta}$ , and  $n$ . Under most situations, you will select either  $\text{Beta}$  or  $n$ .

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

Select  $\text{Beta}$  when you want to calculate the power.

### $n$ (Sample Size)

Enter a value (or range of values) for the sample size,  $n$ . This is the number of individuals sampled in the study. Values must be integers greater than one.

You may enter a range such as 10, 50, 100 or 10 to 100 by 10.

### $N$ (Population Size)

Enter the total number of items in the population from which the sample of  $n$  items is selected. Enter *Infinite* to signify an infinite population so that no correction factor is applied. An *infinite* population is one in which the number in the population is large and unknown.

Note that  $N$  must be greater than  $n$ .

When samples are drawn from a very large (infinite) population, calculations are based on the binomial distribution.

When samples are drawn from a population of known size,  $N$ , calculations are based on the hypergeometric distribution.

### Test Statistic in Report

Specify which test statistic will be used in reporting. Note that *C.C.* is short for *Continuity Correction*. This refers to the adding or subtracting  $1/(2n)$  to (or from) the numerator of the  $z$ -value to bring the normal approximation closer to the binomial distribution.

In most situations, you would select the 'Exact Test' option. The other options are provided for comparative purposes.

## Null Proportion (P0)

Enter a value (or range of values) for the population proportion under the null hypothesis,  $P_0$ . This is the baseline proportion, the proportion that exists in the general population. The proportion estimated from the data will be compared to this value by the statistical test.

Proportions must be between zero and one.

You may enter a range of values such as *0.1,0.2,0.3* or *0.1 to 0.9 by 0.1*.

## Alpha (Significance Level)

This option specifies one or more values for alpha, which is the probability of a type-I (false positive) error. This error occurs when you reject the null hypothesis of equal proportions when in fact they are equal.

Values must be between zero and one. The value of 0.05 is often used for two-sided hypotheses while the value of 0.025 is used for a one-sided hypothesis. You should pick a value for alpha that represents the risk of a type-I you are willing to take in your experimental situation.

Note that because of the discrete nature of the binomial distribution, the alpha level rarely will be achieved exactly.

## Beta (1 - Power)

This option specifies one or more values for beta, which is the probability of a type-II (false negative) error. This error occurs when you fail to reject the null hypothesis of equal proportions when in fact the proportions are different.

Values must be between zero and one. Historically, the value of 0.20 was often used for beta, but now 0.1 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  $1 - \text{beta}$ . 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.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the alternative proportion,  $P1$ .  $P0$  is the value of the proportion assumed by the null hypothesis and  $P1$  (or  $P$ ) is the value of the proportion at which the power is calculated.

### Alternative Hypothesis

This option specifies the alternative hypothesis,  $H1$ . This implicitly specifies the direction of the hypothesis test. Note that the null hypothesis,  $H0$ , is the opposite of  $H1$ .

Note that  $P$  represents the actual value of the proportion and  $P0$  represents the specific value of the proportion assumed by the null hypothesis,  $H0$ .

Possible selections are

**$H1: P \neq P0$** . This is the most common selection. It yields the *two-tailed* test. Use this option when you are testing whether the proportions are different, but you do not want to specify beforehand which proportion is larger. By tradition, most studies are two-tailed unless there is a strong reason to make them one-tailed.

**$H1: P < P0$** . This option yields a *one-tailed* test.

**$H1: P > P0$** . This option also yields a *one-tailed* test.

### Alternative Proportion ( $P1$ )

Enter a value (or range of values) for the value of the binomial proportion at which the power is calculated. This is labeled  $P1$  on the screen. Power calculations assume that this is the true value of the proportion.

This value cannot be equal to  $P0$  since, by definition, it must be an alternative.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P0$ .  $P0$  is the value of the proportion assumed by the null hypothesis and  $P1$  (or  $P$ ) is the value of the proportion at which the power is calculated. Once  $P0$  and the difference are given, the value of  $P1$  is found by the formula:  $P1 = \text{difference} + P0$ .

### Alternative Hypothesis

This option specifies the alternative hypothesis,  $H1$ . This implicitly specifies the direction of the hypothesis test. Note that the null hypothesis,  $H0$ , is the opposite of  $H1$ .

Possible selections are

**H1: Difference  $\neq 0$ .** This is the most common selection. It yields the *two-tailed* test. Use this option when you are testing whether the proportions are different, but you do not want to specify beforehand which proportion is larger. By tradition, most studies are two-tailed unless there is a strong reason to make them one-tailed.

**H1: Difference  $< 0$ .** This option yields a *one-tailed* test.

**H1: Difference  $> 0$ .** This option also yields a *one-tailed* test.

### Alternative Difference (P1-P0)

This option implicitly specifies the value of  $P1$  (the proportion at which the power is calculated) by explicitly specifying the difference. The difference is used with  $P0$  to calculate the value of  $P1$  using the formula,  $P1 = \text{diff} + P0$ .

Since  $P1$  is a proportion, the difference must be between  $-P0$  and  $1 - P0$ . By definition, the difference cannot be zero since  $P1$  is an 'alternative' to  $P0$ .

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P0$ .  $P0$  is the value of the proportion assumed by the null hypothesis and  $P1$  (or  $P$ ) is the value of the proportion at which the power is calculated. Once  $P0$  and the ratio are given, the value of  $P1$  is found by the formula:  $P1 = (P0) \times (\text{ratio})$ .

### Alternative Hypothesis

This option specifies the alternative hypothesis,  $H1$ . This implicitly specifies the direction of the hypothesis test. Note that the null hypothesis,  $H0$ , is the opposite of  $H1$ .

Possible selections are

**H1: Ratio  $\neq 1$ .** This is the most common selection. It yields the *two-tailed* test. Use this option when you are testing whether the proportions are different, but you do not want to specify beforehand which proportion is larger. By tradition, most studies are two-tailed unless there is a strong reason to make them one-tailed.

**H1: Ratio  $< 1$ .** This option yields a *one-tailed* test.

**H1: Ratio  $> 1$ .** This option also yields a *one-tailed* test.

### Alternative Ratio ( $P1/P0$ )

This option implicitly specifies the value of  $P1$  (the proportion at which the power is calculated) by explicitly specifying the ratio. The ratio is used with  $P0$  to calculate the value of  $P1$  using the formula,  $P1 = (P0) \times (\text{ratio})$ .

Since  $P1$  is a proportion, the ratio must be between 0 and 1 / ratio. By definition, the ratio cannot be one since  $P1$  is an 'alternative' to  $P0$ .

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratio,  $O1 / O0$  where  $O1 = P1 / (1 - P1)$  and  $O0 = P0 / (1 - P0)$ .  $P0$  is the value of the proportion assumed by the null hypothesis and  $P1$  (or  $P$ ) is the value of the proportion at which the power is calculated. Once  $P0$  and the odds ratio are given, the value of  $P1$  is found by the formula  $P1 = A / (1 + A)$  where  $A = (O0) \times (\text{odds ratio})$ .

## Alternative Hypothesis

This option specifies the alternative hypothesis,  $H1$ . This implicitly specifies the direction of the hypothesis test. Note that the null hypothesis,  $H0$ , is the opposite of  $H1$ .

Possible selections are

**H1: Odds Ratio  $\neq 1$ .** This is the most common selection. It yields the *two-tailed* test. Use this option when you are testing whether the proportions are different, but you do not want to specify beforehand which proportion is larger. By tradition, most studies are two-tailed unless there is a strong reason to make them one-tailed.

**H1: Odds Ratio  $< 1$ .** This option yields a *one-tailed* test.

**H1: Odds Ratio  $> 1$ .** This option also yields a *one-tailed* test.

## Alternative Odds Ratio ( $O1/O0$ )

This option implicitly specifies the value of  $P1$  (the proportion at which the power is calculated) by explicitly specifying the odds ratio. Since  $P1$  is a proportion, the odds ratio must be greater than zero. By definition, the odds ratio cannot be one since  $P1$  is an 'alternative' to  $P0$ .

## Example1 - Finding the Power

Suppose 50% of patients with a certain type of cancer survive five years. A new treatment will be adopted if it can increase this percentage to 60%. What power will be achieved if a historically controlled trial has sample sizes of 50, 100, 200, 300, 500, or 800 and a significance level of 0.05? For comparative purposes, also calculate the power for alternative proportions of 55% and 65%.

Note: *historically controlled* means that no control group is formed for the current study. Instead, the rates reported from previous studies or that are known to exist in the general population are used. Because of the many advantages that occur when an actual control group is used, historically controlled trials should only be used when a control group is either impossible to obtain or unethical.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the proportion parameterization. You can then make the following entries directly on your screen or you can load the template entitled Example1 by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Proportion (P1).....	<b>0.55 0.60 0.65</b>
Null Proportion (P0) .....	<b>0.50</b>
n .....	<b>50 100 200 300 500 800</b>
N .....	<b>Infinite</b>
Test Statistic in Reports .....	<b>Exact Test</b>
Alternative Hypothesis .....	<b>H1: P &lt;&gt; P0</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>

### Annotated Output

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



## Numeric Results

Numeric Results when  $H_0: P = P_0$  versus  $H_1: P = P < > P_0$   
 Test Statistic: Exact Test

Power	N	Proportion Given $H_0$ ( $P_0$ )	Proportion Given $H_1$ ( $P_1$ )	Target Alpha	Actual Alpha	Beta	Reject $H_0$ If $\leq R   \geq R$
0.0788	50	0.5000	0.5500	0.0500	0.0328	0.9212	17 33
0.1352	100	0.5000	0.5500	0.0500	0.0352	0.8648	39 61
0.2620	200	0.5000	0.5500	0.0500	0.0400	0.7380	85 115
0.3867	300	0.5000	0.5500	0.0500	0.0431	0.6133	132 168
0.5895	500	0.5000	0.5500	0.0500	0.0441	0.4105	227 273
0.7932	800	0.5000	0.5500	0.0500	0.0438	0.2068	371 429
0.2371	50	0.5000	0.6000	0.0500	0.0328	0.7629	17 33
0.4621	100	0.5000	0.6000	0.0500	0.0352	0.5379	39 61
0.7868	200	0.5000	0.6000	0.0500	0.0400	0.2132	85 115
0.9291	300	0.5000	0.6000	0.0500	0.0431	0.0709	132 168
0.9937	500	0.5000	0.6000	0.0500	0.0441	0.0063	227 273
0.9999	800	0.5000	0.6000	0.0500	0.0438	0.0001	371 429
0.5060	50	0.5000	0.6500	0.0500	0.0328	0.4940	17 33
0.8276	100	0.5000	0.6500	0.0500	0.0352	0.1724	39 61
0.9884	200	0.5000	0.6500	0.0500	0.0400	0.0116	85 115
0.9995	300	0.5000	0.6500	0.0500	0.0431	0.0005	132 168
1.0000	500	0.5000	0.6500	0.0500	0.0441	0.0000	227 273
1.0000	800	0.5000	0.6500	0.0500	0.0438	0.0000	371 429

### 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.

$P_0$  is the value of the population proportion under the null hypothesis.

$P_1$  is the value of the population proportion under the alternative hypothesis.

### Summary Statements

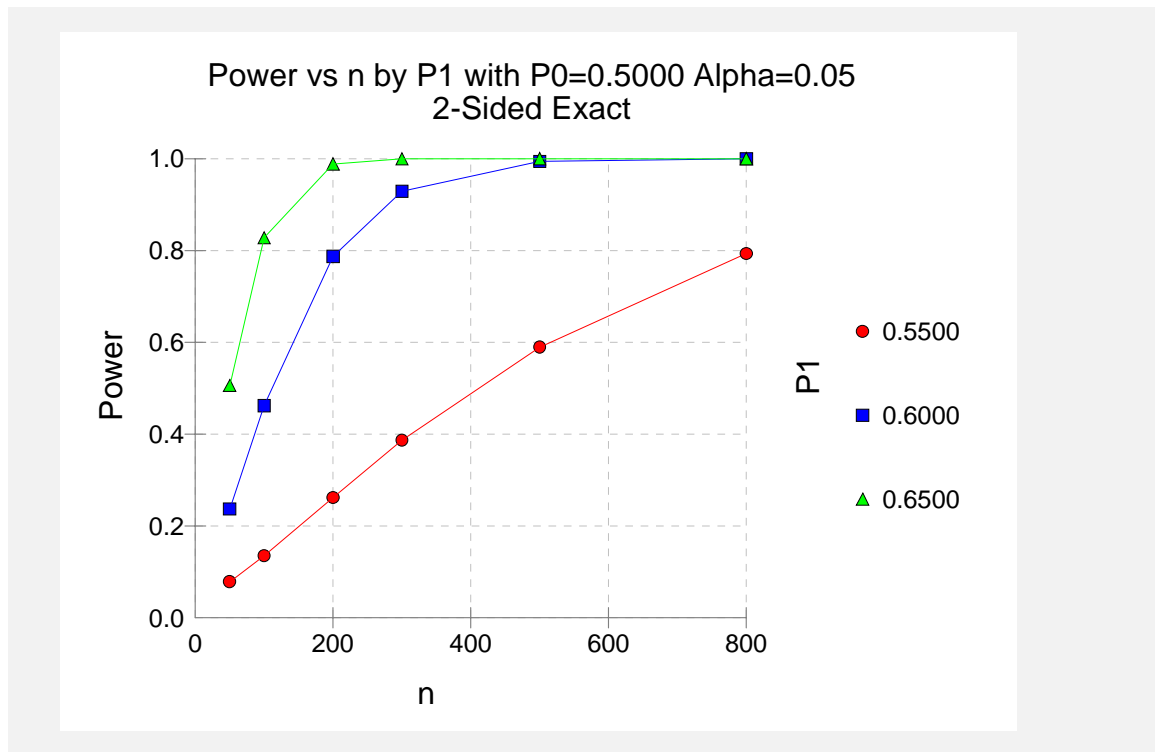
A sample size of 50 achieves 8% power to detect a difference ( $P_1 - P_0$ ) of 0.0500 using a two-sided binomial test. The target significance level is 0.0500. The actual significance level achieved by this test is 0.0328. These results assume that the population proportion under the null hypothesis is 0.5000.

This report shows the values of each of the parameters, one scenario per row. Because of the discrete nature of the binomial distribution, the stated (Target) alpha is usually greater than the actual alpha. Hence, we also show the Actual Alpha along with the rejection region.

The symbol,  $R$ , stands for the number of items with the characteristic of interest out of the  $n$  items sampled. Hence, for the scenario presented on the first line, an exact test does not exist for these parameters at the target alpha of 0.05. The closest that can be achieved is an alpha of 0.0328. In this case, we would reject the null hypothesis in any sample of size 50 in which the count of individuals with the characteristic of interest is less than or equal to 17 or greater than or equal to 33.

The values from this table are plotted in the chart below.

## Plots Section



This plot shows the relationship between power, sample size, and  $P_1$  in this example. We note that 80% power is achieved with a sample size of about 200 when  $P_1$  is 0.60, which was the specific value of interest.

## Example2 - Finding the Sample Size

Continuing with Example1, suppose you want to study the impact of various choices for P1 on sample size. Using a significance level of 0.05 and 90% power, find the sample size when P1 is 0.55, 0.60, 0.65, 0.70, 0.75, and 0.80. Assume that an exact, two-tailed binomial test will be used.

### Setup

First, load the appropriate panel from the menus. This example uses the Proportions parameterization. You can make these changes directly on your screen or you can load the Example2 template by clicking the Template tab and loading this template.

#### Option

#### Value

##### Data Tab

Find .....n  
 Alternative Proportion (P1) .....**0.55 to 0.80 by 0.05**  
 Null Proportion (P0).....**0.50**  
 n .....*Ignored since this is the Find setting*  
 N.....**Infinite**  
 Test Statistic in Reports .....**Exact Test**  
 Alternative Hypothesis .....**H1: P <> P0**  
 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 when H0: P = P0 versus H1: P <>P0

Power	N	Proportion Given H0 (P0)	Proportion Given H1 (P1)	Target Alpha	Actual Alpha	Beta	Reject H0 If <=R >=R
0.9003	1055	0.5000	0.5500	0.0500	0.0487	0.0997	495 560
0.9022	263	0.5000	0.6000	0.0500	0.0483	0.0978	115 148
0.9015	114	0.5000	0.6500	0.0500	0.0487	0.0985	46 68
0.9100	65	0.5000	0.7000	0.0500	0.0464	0.0900	24 41
0.9195	42	0.5000	0.7500	0.0500	0.0436	0.0805	14 28
0.9100	28	0.5000	0.8000	0.0500	0.0357	0.0900	8 20

This report shows the sample sizes corresponding to various values of P1. Notice that a sample size of only 28 is needed to detect the difference between 0.5 and 0.8, but a sample size of 1055 is needed to detect a difference between 0.50 and 0.55.

## Example3 - Investigating the Saw-Tooth Power Function

After releasing the first version of *PASS*, we received many inquiries about the strange shape of the relationship between power and sample size when testing a single proportion using the exact binomial test. This example will show why this strange shape occurs.

### Setup

First, load the appropriate panel. This example uses the Proportions parameterization. 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 Example3 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Null Proportion (P0) .....	<b>0.60</b>
Alternative Proportion (P1).....	<b>0.70</b>
n .....	<b>51 to 60 by 1</b>
N .....	<b>Infinite</b>
Alternative Hypothesis .....	<b>H1: P &lt;&gt; P0</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Reports Tab</b>	
Test Statistic in Reports .....	<b>Exact Test</b>

### Annotated Output

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

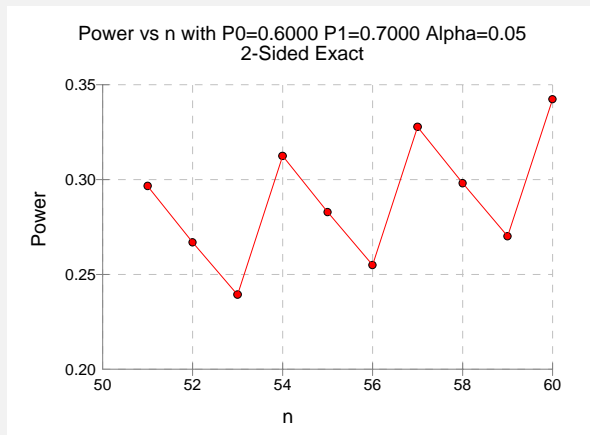
## Numeric Results

Numeric Results for  $H_0: P = P_0$  versus  $H_1: P <> P_0$

Power	N	Proportion Given $H_0$ ( $P_0$ )	Proportion Given $H_1$ ( $P_1$ )	Target Alpha	Actual Alpha	Beta	Reject $H_0$ If $\leq R   \geq R$
0.2966	51	0.6000	0.7000	0.0500	0.0443	0.7034	23 38
0.2669	52	0.6000	0.7000	0.0500	0.0328	0.7331	23 39
0.2393	53	0.6000	0.7000	0.0500	0.0348	0.7607	24 40
0.3124	54	0.6000	0.7000	0.0500	0.0371	0.6876	24 40
0.2828	55	0.6000	0.7000	0.0500	0.0379	0.7172	25 41
0.2549	56	0.6000	0.7000	0.0500	0.0281	0.7451	25 42
0.3277	57	0.6000	0.7000	0.0500	0.0417	0.6723	26 42
0.2981	58	0.6000	0.7000	0.0500	0.0314	0.7019	26 43
0.2701	59	0.6000	0.7000	0.0500	0.0327	0.7299	27 44
0.3423	60	0.6000	0.7000	0.0500	0.0354	0.6577	27 44

This report shows the values of each of the parameters, one scenario per row. The values from this table are plotted in the chart below.

## Plots Section



Notice that the power decreases as  $n$  increases from 51 to 52 and continues to decrease as  $n$  increases to 53. Usually, the power increases as the sample size increases.

To understand why this happens, look at the last column and at the Actual Alpha column. Note that at  $n = 51$ , the actual alpha is 0.0443 and at  $n = 52$ , the actual alpha has decreased to 0.0328. Remember that as alpha decreases, power decreases as well. Hence, increasing the sample size from 51 to 52 was not enough to counterbalance the effect on power of a decrease in alpha from 0.04428 to 0.03281. Hence, the power drops from 0.29656 to 0.26688.

This phenomenon usually occurs for relatively small values of  $n$ .

## Example4 - Step by Step Calculations

In this example, we will take you step by step through the calculations necessary to compute the power of a specific scenario. We will set  $n = 10$ ,  $P_0 = 0.50$ ,  $P_1 = 0.80$ , and  $\alpha = 0.05$ . We will compute the power of the two-tailed test.

### Setup

First, load the appropriate panel from the menus. This example uses the Proportions parameterization. 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 Example4 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Null Proportion (P0) .....	<b>0.8</b>
P0 (Null Proportion) .....	<b>0.5</b>
n .....	<b>10</b>
N .....	<b>Infinite</b>
Test Statistic in Reports .....	<b>Exact Test</b>
Alternative Hypothesis .....	<b>H1: P &lt;&gt; P0</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Reports Tab</b>	
Report Probability Decimals .....	<b>6</b>
Proportion Decimals .....	<b>6</b>

### Annotated Output

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

### Numeric Results

Numeric Results when H0: P = P0 versus H1: P <>P0.							
		Proportion Given H0 (P0)	Proportion Given H1 (P1)	Target Alpha	Actual Alpha	Beta	Reject H0 If <=R >=R
Power	N						
0.375814	10	0.500000	0.800000	0.050000	0.021484	0.624186	1 9

We will now proceed through the calculations necessary to compute this power value.

We first construct a table of binomial probabilities for  $n = 10$  and  $p = 0.5$  using the formula

$$b(r;10,0.5) = \binom{10}{r} 0.5^r (1 - 0.5)^{10-r}$$

**Binomial Probabilities for  $n = 10$  and  $p = 0.5$** 

<b><u>R</u></b>	<b><u>Prob(r = R)</u></b>	<b><u>Cumulative Total</u></b>	<b><u>1 - Cumulative Total</u></b>
0	0.000977	0.000977	0.999023
1	0.009766	0.010742	0.989258
2	0.043945	0.054688	0.945313
3	0.117188	0.171875	0.828125
4	0.205078	0.376953	0.623047
5	0.246094	0.623047	0.376953
6	0.205078	0.828125	0.171875
7	0.117188	0.945313	0.054688
8	0.043945	0.989258	0.010742
9	0.009766	0.999023	0.000977
10	0.000977	1.000000	0.000000

When we construct a two-tailed test, we split the alpha value evenly between the two tails. Hence, we place  $\alpha / 2$  (or 0.025) in each tail. Moving down from the top, we find that the cumulative probability is 0.010742 for  $R = 1$  and 0.054688 for  $R = 2$ . Since 0.054688 is greater than 0.025, we adopt  $R = 1$  as our lower rejection value. Likewise, we find that  $R = 9$  is the upper rejection value.

Our testing strategy is

1. Draw a sample of 10 items and count the number with the characteristic of interest. Call this value  $r$ .
2. If  $r = 0, 1, 9$ , or  $10$ , reject the null hypothesis that  $p = 0.5$  in favor of the alternative hypothesis that  $p \neq 0.5$ .

Now, to compute the power for  $P1 = 0.8$ , we must compute another table of binomial probabilities, this time for  $p = 0.8$  using the formula.

$$b(r;10,0.8) = \binom{10}{r} 0.8^r (1 - 0.8)^{10-r}$$

**Binomial Probabilities for  $n = 10$  and  $p = 0.8$** 

<b><u>R</u></b>	<b><u>Prob(r = R)</u></b>	<b><u>Cumulative Total</u></b>	<b><u>1 - Cumulative Total</u></b>
0	0.000000	0.000000	1.000000
1	0.000004	0.000004	0.999996
2	0.000074	0.000078	0.999922
3	0.000786	0.000864	0.999136
4	0.005505	0.006369	0.993631
5	0.026424	0.032793	0.967207
6	0.088080	0.120874	0.879126
7	0.201327	0.322200	0.677800
8	0.301990	0.624190	0.375810
9	0.268435	0.892626	0.107374
10	0.107374	1.000000	0.000000

The power is the probability of rejecting the null hypothesis. This occurs when  $r = 0, 1, 9$ , or  $10$ . From the above table, we compute the power as  $0.000000 + 0.000004 + 0.268435 + 0.107374 = 0.375813$ . This matches the calculated power value as displayed in the results above to within rounding error.

## Example5 - Validation using Zar

Zar (1984) page 388 gives the results of a power analysis. When  $n = 12$ ,  $P_0 = 0.50$ ,  $P_1 = 0.83$ , and  $\alpha = 0.05$  using a one-sided test, Zar reports a power of 0.666.

### Setup

First, load the appropriate panel from the menus. This example uses the Proportions parameterization. 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 Example5 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Proportion (P1).....	<b>0.83</b>
Null Proportion (P0) .....	<b>0.5</b>
n .....	<b>12</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
N .....	<b>Infinite</b>
Alternative Hypothesis .....	<b>H1: P&gt;P0</b>

### Annotated Output

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

### Numeric Results

Numeric Results for H0: P = P0 versus Ha: P = P1>P0							
Power	N	Proportion Given H0 (P0)	Proportion Given H1 (P1)	Target Alpha	Actual Alpha	Beta	Reject H0 If R>=This
0.6656	12	0.5000	0.8300	0.0500	0.0193	0.3344	10

*PASS* calculated the power as 0.6656, which agrees with Zar's value of 0.666.



## Example6 – Comparing Test Statistics

One important decision that must be made before conducting the experiment is to decide which of the available test statistics to use. This procedure makes it easy to make this comparison. The parameter settings will be set as they were in Example 1 except that the alternative proportion is set to 0.60 and the sample sizes are 10, 11, 12, 25, 50, and 70.

### Setup

First, load the appropriate panel from the menus. This example uses the Proportions parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Hypothesis .....	<b>H1: <math>P \neq P_0</math></b>
Alternative Proportion (P1) .....	<b>0.60</b>
Null Proportion (P0) .....	<b>0.50</b>
n .....	<b>10 11 12 25 50 70</b>
N .....	<b>Infinite</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Reports Tab</b>	
Show Comparative Reports .....	<b>Checked</b>
Show Comparative Plots .....	<b>Checked</b>

### Annotated Output

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

## Numeric Results

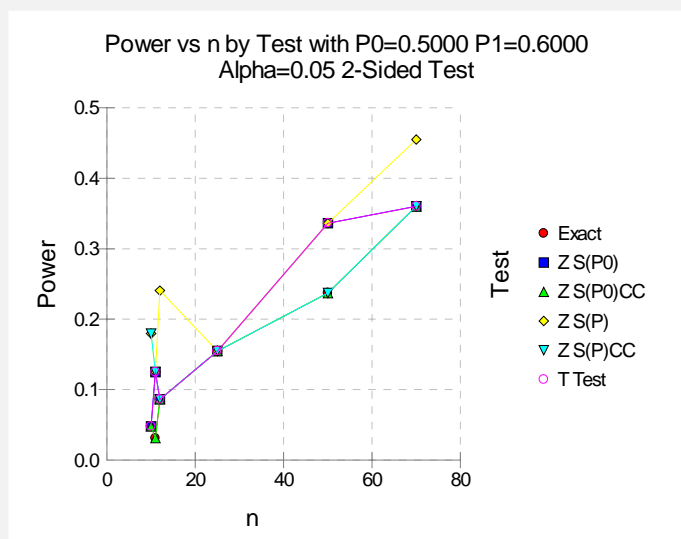
Power Comparison for Methods of Testing  $H_0: P = P_0$  versus  $H_1: P \neq P_0$

N	P0	P1	Alpha	Exact Test Power	Z-Test S(P0) Power	Z-Test S(P0)C Power	Z-Test S(P) Power	Z-Test S(P)C Power	T-Test Power
10	0.5000	0.6000	0.0500	0.0480	0.0480	0.0480	0.1796	0.1796	0.0480
11	0.5000	0.6000	0.0500	0.0310	0.1248	0.0310	0.1248	0.1248	0.1248
12	0.5000	0.6000	0.0500	0.0863	0.0863	0.0863	0.2406	0.0863	0.0863
25	0.5000	0.6000	0.0500	0.1548	0.1548	0.1548	0.1548	0.1548	0.1548
50	0.5000	0.6000	0.0500	0.2371	0.3361	0.2371	0.3361	0.2371	0.3361
70	0.5000	0.6000	0.0500	0.3601	0.3601	0.3601	0.4549	0.3601	0.3601

Actual Alpha Comparison for Methods of Testing  $H_0: P = P_0$  versus  $H_1: P \neq P_0$

N	P0	P1	Target Alpha	Exact Test Alpha	Z-Test S(P0) Alpha	Z-Test S(P0)C Alpha	Z-Test S(P) Alpha	Z-Test S(P)C Alpha	T-Test Alpha
10	0.5000	0.6000	0.0500	0.0215	0.0215	0.0215	0.1094	0.1094	0.0215
11	0.5000	0.6000	0.0500	0.0117	0.0654	0.0117	0.0654	0.0654	0.0654
12	0.5000	0.6000	0.0500	0.0386	0.0386	0.0386	0.1460	0.0386	0.0386
25	0.5000	0.6000	0.0500	0.0433	0.0433	0.0433	0.0433	0.0433	0.0433
50	0.5000	0.6000	0.0500	0.0328	0.0649	0.0328	0.0649	0.0328	0.0649
70	0.5000	0.6000	0.0500	0.0414	0.0414	0.0414	0.0722	0.0414	0.0414

Chart Section



An examination of the first report shows that for most sample sizes, the power is different for at least one of the tests. Also, notice that the exact test always has the minimum power in each row. This would lead us to discard this test statistic. However, consider the second report, which shows the actual alpha level (the target was 0.05) for each test. By inspecting corresponding entries in both tables, we can see that whenever a test statistic achieves a better power than the exact test, it also exceeds the target alpha. For example, look at the powers for  $n = 12$ . The z test using  $s(\hat{p})$  has an unusually large power of 0.2406. This is a much larger power than the exact test's value of 0.0863. However, note that the actual alpha level for this test is 0.1460, which is much higher than the target of 0.05 and the actual value of the other tests, which is 0.0386.

We conclude that indeed, the exact test is consistently the best test since it always achieves a significance level that is less than the target value.

## Example7 - Finding the Power using Ratios

Suppose that only 5% of patients with an aggressive type of cancer respond to the standard treatment. Researchers have found a new treatment which could be widely used if the percentage of patients responding is at least 0.5 times greater than the proportion responding to the standard treatment, i.e.  $P1 = 1.5(P0)$ , or in terms of ratios,  $P1/P0 = 1.5$ . What power will be achieved for trials with sample sizes of 200, 300, 500, or 800 and a significance level of 0.05? For comparative purposes, also calculate the power for alternative ratios of 1.25 and 1.75.

### Setup

First, load the appropriate panel from the menus. This example uses the Ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Ratios (P1/P0) .....	<b>1.25 1.50 1.75</b>
Null Proportion (P0) .....	<b>0.05</b>
n .....	<b>200 300 500 800</b>
N .....	<b>Infinite</b>
Test Statistic in Reports .....	<b>Exact Test</b>
Alternative Hypothesis .....	<b>H1: Ratio <math>\neq</math> 1</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>

### Annotated Output

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

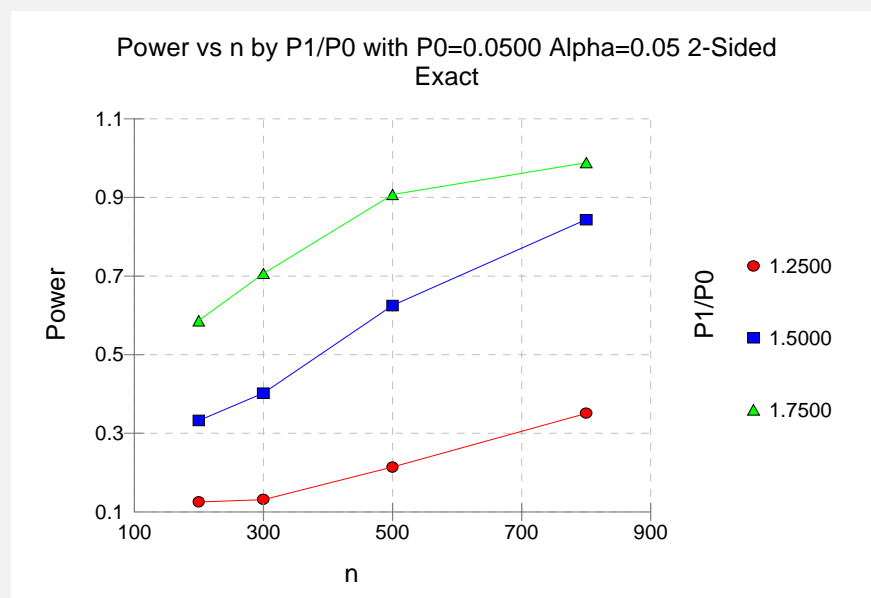
## Numeric Results

Numeric Results for testing  $H_0: P = P_0$  versus  $H_1: P \neq P_0$   
 Test Statistic: Exact Test

Power	N	Proportion Given $H_0$ ( $P_0$ )	Proportion Given $H_1$ ( $P_1$ )	Ratio ( $P_1 / P_0$ )	Target Alpha	Actual Alpha	Beta	Reject $H_0$ If $\leq R$   $\geq R$
0.1247	200	0.0500	0.0625	1.2500	0.0500	0.0328	0.8753	3 17
0.1315	300	0.0500	0.0625	1.2500	0.0500	0.0328	0.8685	7 24
0.2138	500	0.0500	0.0625	1.2500	0.0500	0.0395	0.7862	15 36
0.3509	800	0.0500	0.0625	1.2500	0.0500	0.0420	0.6491	27 53
0.3322	200	0.0500	0.0750	1.5000	0.0500	0.0328	0.6678	3 17
0.4019	300	0.0500	0.0750	1.5000	0.0500	0.0328	0.5981	7 24
0.6248	500	0.0500	0.0750	1.5000	0.0500	0.0395	0.3752	15 36
0.8432	800	0.0500	0.0750	1.5000	0.0500	0.0420	0.1568	27 53
0.5861	200	0.0500	0.0875	1.7500	0.0500	0.0328	0.4139	3 17
0.7062	300	0.0500	0.0875	1.7500	0.0500	0.0328	0.2938	7 24
0.9072	500	0.0500	0.0875	1.7500	0.0500	0.0395	0.0928	15 36
0.9882	800	0.0500	0.0875	1.7500	0.0500	0.0420	0.0118	27 53

This report shows the values of each of the parameters, one scenario per row. Because of the discrete nature of the binomial distribution, the stated (Target) alpha is usually greater than the actual alpha. Hence, we also show the Actual Alpha along with the rejection region.

The values from this table are plotted in the chart below.



This plot shows the relationship between power and  $P_1/P_0$  in this example. We note that 80% power is achieved with a sample size of about 720 when  $P_1/P_0$  is 1.50, which was the specific ratio of interest.

## Example8 – Determining the Power after Completing an Experiment

A group of researchers is studying the effects of a new diet on cholesterol levels in high-risk patients. The researchers had hypothesized that the cholesterol level would be reduced to a safe level in more than 70% of subjects following the new diet. They are confident that the proportion will be no less than 0.70. To test this one-sided hypothesis, they randomly sampled 200 individuals with dangerously high cholesterol and put them on the new diet. After the period of the study, the researchers determined that 150 of the 200 patients sampled (75%) had reduced their cholesterol level while on the new diet. Statistical analysis using the exact test and an alpha level of 0.05, however, resulted in failure to reject the null hypothesis that the proportion is 0.70. The researchers desire now to compute the power of their study for true proportions ranging from 0.71 to 0.80.

Note that a range of proportions is considered for power calculations instead of just 0.75, the sample proportion found in the experiment. While it is tempting to use the sample proportion as the true proportion in post-experiment power calculations, it is more informative to review a range of possible alternatives representing practically significant differences from the null value.

### Setup

First, load the appropriate panel from the menus. This example uses the Proportions parameterization. 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.

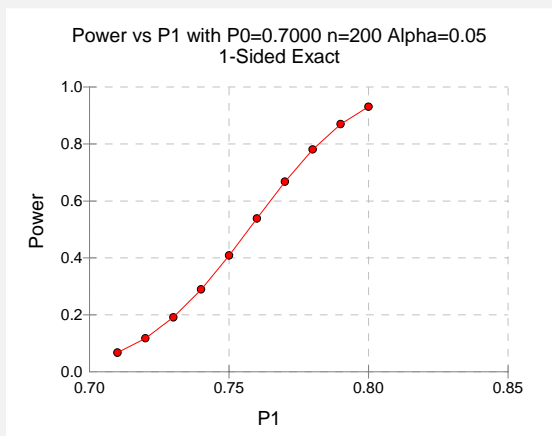
<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Proportion (P1) .....	<b>0.71 to 0.80 by 0.01</b>
Null Proportion (P0) .....	<b>0.70</b>
n .....	<b>200</b>
N .....	<b>Infinite</b>
Test Statistic in Reports .....	<b>Exact Test</b>
Alternative Hypothesis .....	<b>H1: P &gt; P0</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>

## Annotated Output

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

**Numeric Results for testing  $H_0: P = P_0$  versus  $H_1: P > P_0$**   
**Test Statistic: Exact Test**

Power	N	Proportion Given $H_0$ ( $P_0$ )	Proportion Given $H_1$ ( $P_1$ )	Target Alpha	Actual Alpha	Beta	Reject $H_0$ If $R \geq \text{This}$
0.0675	200	0.7000	0.7100	0.0500	0.0359	0.9325	152
0.1178	200	0.7000	0.7200	0.0500	0.0359	0.8822	152
0.1913	200	0.7000	0.7300	0.0500	0.0359	0.8087	152
0.2894	200	0.7000	0.7400	0.0500	0.0359	0.7106	152
0.4083	200	0.7000	0.7500	0.0500	0.0359	0.5917	152
0.5386	200	0.7000	0.7600	0.0500	0.0359	0.4614	152
0.6673	200	0.7000	0.7700	0.0500	0.0359	0.3327	152
0.7807	200	0.7000	0.7800	0.0500	0.0359	0.2193	152
0.8696	200	0.7000	0.7900	0.0500	0.0359	0.1304	152
0.9310	200	0.7000	0.8000	0.0500	0.0359	0.0690	152



Power ranges from 0.0675 for a true proportion of 0.71 to 0.9310 for a true proportion of 0.80.

## Chapter 105

# One Proportion – Non-Inferiority

This module provides power analysis and sample size calculation for non-inferiority and superiority tests in one-sample designs in which the outcome is binary. Users may choose from among six popular test statistics commonly used for running the hypothesis test.

The details of sample size calculation for the one-sample design for binary outcomes are presented in the chapter One Proportion - Equality and they will not be duplicated here. Instead, this chapter focuses on those changes necessary for non-inferiority and superiority tests.

Approximate sample size formulas for non-inferiority tests of a single proportion are presented in Chow et al. (2003) page 83. However, only large sample (normal approximation) results are given there. The results available in this module use exact calculations based on the enumeration of all possible values of the binomial distribution.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Example

A non-inferiority test example will set the stage for the discussion of the terminology that follows. Suppose that the current treatment for a disease is effective 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the first questions that must be answered is whether the new treatment is as good as the current treatment. In other words, do at least 70% of treated subjects respond to the new treatment?

Because of the many benefits of the new treatment, clinicians are willing to adopt the new treatment even if it is slightly less effective than the current treatment. They must determine, however, how much less effective the new treatment can be to still be adopted. Should it be adopted if 69% respond? 68%? 65%? 60%? There is a percentage below 70% at which the difference between the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, it was decided that if a response of at least 63% was achieved, the new treatment would be adopted. The difference between these two percentages is called the *margin of equivalence*. The margin of equivalence in this example is 7%.

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is at least 0.63. The statistical hypothesis to be tested is

$$H_0: P \leq 0.63 \text{ versus } H_1: P > 0.63$$

Notice that when the null hypothesis is rejected, the conclusion is that the response rate is at least 0.63. Note that even though the response rate of the current treatment is 0.70, the hypothesis test is about a response rate of 0.63. Also notice that a rejection of the null hypothesis results in the conclusion of interest.



## Technical Details

In the discussion that follows, let  $P$  represent the proportion responding as a success. That is,  $P$  is the actual probability of a *success* in a binomial experiment. Let  $PB$  represent the *baseline* proportion. In a non-inferiority experiment, the baseline proportion is the response rate of the current treatment. Furthermore, let  $P_0$  represent the response proportion that is tested in the null hypothesis,  $H_0$ . The power of a test is computed at a specific value of the proportion. Let  $P_1$  represent the proportion at which the power is computed.

Let  $PE$  represent the smallest value of  $P$  that still results in the conclusion that the new treatment is noninferior to the current treatment. The statistical hypotheses that are tested are

$$H_0: P \leq PE \text{ versus } H_1: P > PE$$

There are three common methods of specifying the margin of equivalence. The most direct is to simply assign values for  $PB$  and  $PE$ . However, it is often more meaningful to identify  $PB$  and then specify  $PE$  implicitly by giving their difference, ratio, or odds ratio. Mathematically, the definitions of these parameterizations are

<u>Parameter</u>	<u>Computation</u>	<u>Hypotheses</u>
Difference	$d_0 = PE - PB$	$H_0: P \leq PB + d_0$ vs $H_1: P > PB + d_0$
Ratio	$r_0 = PE / PB$	$H_0: P \leq r_0(PB)$ vs $H_1: P > r_0(PB)$
Odds Ratio	$o_0 = OddsE / OddsB$	$H_0: P \leq A$ vs $H_1: P > A$ where $A = \frac{o_0(PB)}{1 + PB(o_0 - 1)}$

## Difference

The difference is perhaps the most direct method of comparison between two proportions. It is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One difficulty arises when the event of interest is rare. If a difference of 0.001 occurs when the baseline probability is 0.40, it would be dismissed as being trivial. That is, there is usually little interest in a treatment that only decreases the probability from 0.400 to 0.399. However, if the baseline probability of a disease is 0.002, a 0.001 decrease would represent a reduction of 50%. Thus, interpretation of the difference depends on the baseline probability of the event. As a rule of thumb, the difference is best suited for those cases in which  $0.20 < P < 0.80$ .

Note that if  $d_0 < 0$ , the procedure is called a *non-inferiority test* while if  $d_0 > 0$  the procedure is called a *superiority test*.

## Non-Inferiority using a Difference

The following example might help you understand the concept of a *non-inferiority test*. Suppose 60% of patients respond to the current treatment method ( $P_B = 0.60$ ). If the response rate of the new treatment is no less than five percentage points worse ( $d_0 = -0.05$ ) than the existing treatment, it will be considered noninferior. Substituting these figures into the statistical hypotheses gives  $H_0: d \leq -0.05$  versus  $H_1: d > -0.05$ . The relationship  $P_0 = P_B + d_0$

gives  $H_0: P \leq 0.55$  versus  $H_1: P > 0.55$ .

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 55%.

## Superiority using a Difference

The following example is intended to help you understand the concept of *superiority*. Suppose 60% of patients respond to the current treatment method ( $P_B = 0.60$ ). If the response rate of the new treatment is at least ten percentage points better ( $d_0 = 0.10$ ), it will be considered to be superior to the existing treatment. Substituting these figures into the statistical hypotheses gives  $H_0: d \leq 0.10$  versus  $H_1: d > 0.10$ . The relationship  $P_0 = P_B + d_0$  gives  $H_0: P \leq 0.70$  versus  $H_1: P > 0.70$ .

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 0.70. That is, the conclusion of superiority is that the new treatment's response rate is at least 0.10 more than that of the existing treatment.

## Ratio

The ratio  $r_0 = PE / PB$  gives the relative change in the probability of the response. Testing non-inferiority and superiority use the same formulation  $H_0: r \leq r_0$  versus  $H_1: r > r_0$ .

The only subtlety is that for non-inferiority tests  $r_0 < 1$  while for superiority tests  $r_0 > 1$ .

### Non-Inferiority using a Ratio

The following example might help you understand the concept of *non-inferiority* as defined by the ratio. Suppose that 60% of patients ( $PB = 0.60$ ) respond to the current treatment method. If a new treatment decreases the response rate by no more than 10% ( $r_0 = 0.90$ ), it will be considered to be noninferior to the standard treatment. Substituting these figures into the statistical hypotheses gives  $H_0: r \leq 0.90$  versus  $H_1: r > 0.90$ . The relationship  $P_0 = (r_0)(PB)$  gives  $H_0: P \leq 0.54$  versus  $H_1: P > 0.54$ .

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 54%. That is, the conclusion of non-inferiority is that the new treatment's response rate is no worse than 10% less than that of the standard treatment.

### Superiority using a Ratio

The following example is intended to help you understand the concept of *superiority* as it applies to the ratio. Suppose that 60% of patients ( $PB = 0.60$ ) respond to the current treatment method. If a new treatment increases the response rate by at least 10% ( $r_0 = 1.10$ ), it will be considered to be superior to the existing treatment. Substituting these figures into the statistical hypotheses gives  $H_0: r \leq 1.10$  versus  $H_1: r > 1.10$ . The relationship  $P_0 = (r_0)(PB)$  gives  $H_0: P \leq 0.66$  versus  $H_1: P > 0.66$ .

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 66%. That is, the conclusion of superiority is that the new treatment's response rate is at least 10% more than that of the existing treatment.

## Odds Ratio

The odds ratio,  $o0 = (PE / (1 - PE)) / (PB / (1 - PB))$ , gives the relative change in the odds of the response. Testing non-inferiority and superiority use the same formulation  $H0: o \leq o0$  versus  $H1: o > o0$ . The only difference is that for non-inferiority tests  $o0 < 1$ , while for superiority tests  $o0 > 1$ .

## Power and Sample Size Calculation

Historically, power and sample size calculations for a one-sample proportion test have been based on normal approximations to the binomial. However, with the speed of modern computers using the normal approximation is unnecessary, especially for small samples. Rather, the significance level and power can be computed using complete enumeration of all possible values of  $x$ , the number of successes in a sample of size  $n$ .

This is done as follows.

1. The critical value of the test is computed using standard techniques.
2. For each possible value of  $x$ , the value of the test statistic ( $z$  test,  $t$  test, or exact test) is computed along with its associated probability of occurrence.
3. The significance level and power are computed by summing the probabilities of occurrence for all values of the test statistic that are greater than (or less than) the critical value. Each probability of occurrence is calculated using  $P0$  for the significance level and  $P1$  for the power.

Other variables such as the sample size are then found using an efficient search algorithm. Although this method is not as elegant as a closed-form solution, it is completely accurate.

## Test Statistics

The test statistics used are listed in the One Proportion - Inequality chapter. They will not be repeated 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *Equiv. Value*, *Actual Value*, *Alpha*, *Beta*, and *n*. In most situations, you will select either Beta or *n*.

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.

### Higher Proportions Are

This option specifies whether proportions represent successes or failures.

#### Better

When proportions represent successes, the higher proportions are better. In this case, a non-inferior treatment is one whose proportion is at least as high as the baseline. The alternative hypothesis of non-inferiority is  $H_1: P > P_0$ , where  $P_0$  is slightly less than PB.

#### Worse

When proportions represent failures, the lower proportions are better. In this case, a noninferior treatment is one whose proportion is at least as low as the baseline or lower. The alternative hypothesis of non-inferiority is  $H_1: P < P_0$ , where  $P_0$  is slightly greater than PB.

### Baseline Proportion (PB)

Enter a value (or range of values) for the baseline proportion. In a non-inferiority study, this is the response rate of the standard (existing) treatment. Note that this is not the value of  $P_0$ . Instead, this value is used in the calculation of  $P_0$ .

Proportions must be between zero and one.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

### n (Sample Size)

Enter a value (or range of values) for the sample size *n*. This is the number of individuals sampled in the study. Values must be integers greater than one.

You may enter a range such as *10, 50, 100* or *10 to 100 by 10*.

## N (Population Size)

Enter the total number of items in the population from which the sample of  $n$  items is selected. Enter *Infinite* to signify an infinite population so that no correction factor is applied. An *infinite* population is one in which the number in the population is large and unknown.

Note that  $N$  must be greater than  $n$ .

When samples are drawn from a very large (infinite) population, calculations are based on the binomial distribution.

When samples are drawn from a population of known size  $N$ , calculations are based on the hypergeometric distribution.

## Alpha (Significance Level)

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

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. You should pick a value for alpha that represents the risk of a type-I you are willing to take in your experimental situation.

Note that because of the discrete nature of the binomial distribution, the alpha level will seldom be achieved exactly.

## Beta (1 - Power)

This option specifies one or more values for beta which is the probability of a type-II (false negative) error. This 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, but now 0.1 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. 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.

## Test Statistic in Report

Specify which test statistic will be used in searching and reporting. Note that *C.C.* is an abbreviation for *Continuity Correction*. This refers to the adding or subtracting of  $1/(2n)$  to (or from) the numerator of the z-value to bring the normal approximation closer to the binomial distribution.

In most situations, you would select the Exact Test option. The other options are provided for comparative purposes.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is given directly in terms of the proportions  $P_0$  and  $P_1$ .

### Equivalence Proportion ( $P_0$ )

This option sets the smallest value which is still trivially different from  $P_B$  by directly setting the value of  $P_0$ . If 'Higher Proportions Are' is set to 'Better', specify a value of  $P_0$  that is less than  $P_B$  for a non-inferiority test or a value of  $P_0$  that is greater than  $P_B$  for a superiority test. If 'Higher Proportions Are' is set to 'Worse', do the opposite.

For example, if  $P_B$  (baseline proportion) is 0.50, you might consider 0.49, 0.48, 0.47, and 0.46 to be close enough so that the fact that they are less than 0.50 can be overlooked. However, you might decide that if the value is 0.45 or less, the treatment is inferior. Thus, this value would be set to 0.45.

Since this value is a proportion, it must be a positive value less than one. It cannot be equal to  $P_B$ .

### Actual Proportion ( $P_1$ )

This is the value of the proportion ( $P_1$ ) at which the power is calculated. The power calculations assume that this is the actual value of the proportion. For non-inferiority tests, this value is often set equal to  $P_B$ .

Proportions must be between zero and one.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P0$ .  $P0$  is the value of the proportion assumed by the null hypothesis and  $P1$  (or  $P$ ) is the value of the proportion at which the power is calculated. Once  $P0$  and the difference are given, the value of  $P1$  is found by the formula:  $P1 = \text{difference} + P0$ .

### Equivalence Difference (d0)

This option sets the smallest value which is still trivially different from  $PB$  by setting the difference between  $P0$  and  $PB$ . If 'Higher Proportions Are' is set to 'Better', specify a difference that is less than zero for a non-inferiority test or a difference greater than zero for a superiority test. If 'Higher Proportions Are' is set to 'Worse', do the opposite.

For example, if  $PB$  (baseline proportion) is 0.50, you might consider -.01, -.02, or -.04 to be small enough so that the fact that  $P0$  is less than 0.50 can be overlooked. However, you might decide that if the difference is -.05 or less, the treatment is inferior. Thus, this value would be set to -.05.

Since this value is a difference between two proportions, it must be between -1 and 1.

### Actual Difference (d1)

This option specifies the value of  $P1$  (the proportion at which the power is calculated) by specifying the difference between the two proportions,  $P1$  and  $PB$ . This difference is used with  $PB$  to calculate the value of  $P1$  using the formula,  $P1 = PB + \text{Difference}$ . For non-inferiority tests, this value is often set equal to zero.

Differences must be between -1 and 1.

You may enter a range of values such as *.03 .05 .10* or *.01 to .05 by .01*.



## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P0$ .

### Equivalence Ratio (r0)

This option sets the smallest value which is still trivially different from PB by setting the ratio of P0 to PB. If 'Higher Proportions Are' is set to 'Better', specify a ratio that is less than one for a non-inferiority test or a ratio greater than one for a superiority test. If 'Higher Proportions Are' is set to 'Worse', do the opposite.

For example, if PB (baseline proportion) is 0.50, you might consider ratios of 0.99, 0.98, or even 0.96 to be small enough so that the fact that P0 is less than PB can be overlooked (the difference is trivial). However, you might decide that if the ratio is 0.95 or less, the treatment is inferior. Thus, this value would be set to 0.95.

Since this value is a ratio between two proportions, it must be positive. Since it is a margin, it cannot be one. It cannot be so large that the calculated value of P0 is greater than one.

### Actual Ratio (r1)

This option specifies the value of P1 (the actual proportion) by specifying the ratio between the two proportions, P1 and PB. This ratio is used with PB to calculate the value of P1 using the formula,  $P1 = PB(\text{Ratio})$ . For non-inferiority tests, this value is often set equal to one.

Ratios must greater than zero. Note that the ratios must be small enough so that P1 is less than one.

You may enter a range of values such as *.5 .6 .7 .8* or *1.25 to 2.0 by .25*.

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratio,  $O1 / O0$  where  $O1 = P1 / (1 - P1)$  and  $O0 = P0 / (1 - P0)$ .

### Equivalence Odds Ratio (o0)

This option sets the smallest value which is still trivially different from PB by setting the odds ratio of P0 and PB. If 'Higher Proportions Are' is set to 'Better', specify a ratio that is less than one for a non-inferiority test or a ratio greater than one for a superiority test. If 'Higher Proportions Are' is set to 'Worse', do the opposite.

For example, if PB (baseline proportion) is 0.50, you might consider odds ratios of 0.99, 0.90, or even 0.81 to be small enough so that the fact that P0 is less than PB can be overlooked (the difference is trivial). However, you might decide that if the odds ratio is 0.80 or less, the treatment is inferior. Thus, this value would be set to 0.80.

Since this value is a ratio between two odds, it must be positive. Since it is a margin, it cannot be one.

### Actual Odds Ratio (o1)

This option specifies the value of P1 (the actual proportion) by specifying the odds ratio between the two proportions, P1 and PB. This ratio is used with PB to calculate the value of P1. For non-inferiority tests, this value is often set equal to one.

Odds ratios must greater than zero.

You may enter a range of values such as *.5 .6 .7 .8* or *1.25 to 2.0 by .25*.

## Example 1 - Finding the Power

Suppose 50% of patients with a certain type of cancer survive two years using the current treatment. The current treatment is expensive and has several severe side effects. A new treatment has fewer side effects and is less expensive. A non-inferiority trial is to be conducted to show that the two-year survival rate of the new treatment is as good as the current treatment. After serious consideration, the margin of non-inferiority is set at 5%. What power will be achieved by sample sizes of 50, 100, 200, 300, 500, or 800 and a significance level of 0.05? For comparative purposes, also calculate the power for a margin of non-inferiority of 10%. Assume that the true survival rate of the new treatment is the same as that of the current (baseline) treatment.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. You can make these changes directly on your screen or you can load the template Example1 by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Higher Proportions Are.....	<b>Better</b>
Equivalence Difference (d0) .....	<b>-.10 -.05</b>
Actual Difference (d1) .....	<b>0</b>
Baseline Proportion (PB).....	<b>0.50</b>
n (Sample Size).....	<b>50 100 200 300 500 800</b>
N (Population Size) .....	<b>Infinite</b>
Alpha .....	<b>0.05</b>
Beta (1-Power) .....	<i>Ignored since this is the Find setting</i>
Test Statistic in Reports .....	<b>Exact Test</b>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>
Show Comparative Reports .....	<b>Not checked</b>
Show Definitions .....	<b>Checked</b>
Show Plots .....	<b>Checked</b>
Show Comparative Plots.....	<b>Not checked</b>

### Annotated Output

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

## Numeric Results

Numeric Results when  $H_0: P = P_0$  versus  $H_1: P = P > P_0$   
 Test Statistic: Exact Test

Power	N	Equiv. Difference (d0)	Actual Difference (d1)	Baseline Proportion (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 If $R \geq \text{This}$
0.3359	50	-0.1000	0.0000	0.5000	0.0500	0.0314	0.6641	27
0.6178	100	-0.1000	0.0000	0.5000	0.0500	0.0423	0.3822	49
0.8854	200	-0.1000	0.0000	0.5000	0.0500	0.0492	0.1146	92
0.9633	300	-0.1000	0.0000	0.5000	0.0500	0.0443	0.0367	135
0.9976	500	-0.1000	0.0000	0.5000	0.0500	0.0461	0.0024	219
1.0000	800	-0.1000	0.0000	0.5000	0.0500	0.0453	0.0000	344
0.1611	50	-0.0500	0.0000	0.5000	0.0500	0.0444	0.8389	29
0.2421	100	-0.0500	0.0000	0.5000	0.0500	0.0441	0.7579	54
0.3619	200	-0.0500	0.0000	0.5000	0.0500	0.0381	0.6381	103
0.5230	300	-0.0500	0.0000	0.5000	0.0500	0.0465	0.4770	150
0.7195	500	-0.0500	0.0000	0.5000	0.0500	0.0484	0.2805	244
0.8783	800	-0.0500	0.0000	0.5000	0.0500	0.0476	0.1217	384

### 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.

Equiv. is the maximum value that is still considered unimportant.

Actual is the value of this parameter given the alternative hypothesis is true.

PB is the baseline or standard value of the proportion. This is the value under the current treatment.

d0 is the smallest difference from PB which is still considered as equivalent.

d1 is the value of the difference under the alternative hypothesis.

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 50 achieves 34% power to detect a difference ( $P_0 - PB$ ) of -0.1000 using a one-sided binomial test. The target significance level is 0.0500. The actual significance level achieved by this test is 0.0314. These results assume a baseline proportion (PB) of 0.5000 and that the actual difference ( $P_1 - PB$ ) is 0.0000.

This report shows the values of each of the parameters, one scenario per row. Because of the discrete nature of the binomial distribution, the target alpha is usually greater than the actual alpha. Hence, the actual alpha is also shown.

## Power

Power is the probability of concluding non-inferiority when the treatment is indeed noninferior.

## N

This is the sample size.

## Equiv. Difference (or Proportion, Ratio, or Odds Ratio)

This difference is the maximum difference from the baseline proportion PB that is still considered unimportant or trivial. This value is used to calculate  $P_0$ .

## Actual Difference (or Proportion, Ratio, or Odds Ratio)

The actual difference is difference between the true proportion,  $P_1$ , and the baseline proportion, PB.

## Baseline Proportion

The baseline proportion is the response rate that is achieved by the current (standard) treatment.

## Target Alpha

This is the target (set in the design) value of the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. That is, this is the probability of concluding non-inferiority when in fact the new treatment is inferior. Because of the discreteness of the binomial distribution from which this value is calculated, the target value is seldom achieved exactly.

## Actual Alpha

This is the actual value of alpha (see Target Alpha) that is achieved by the design. Note that lower values of alpha imply lower power.

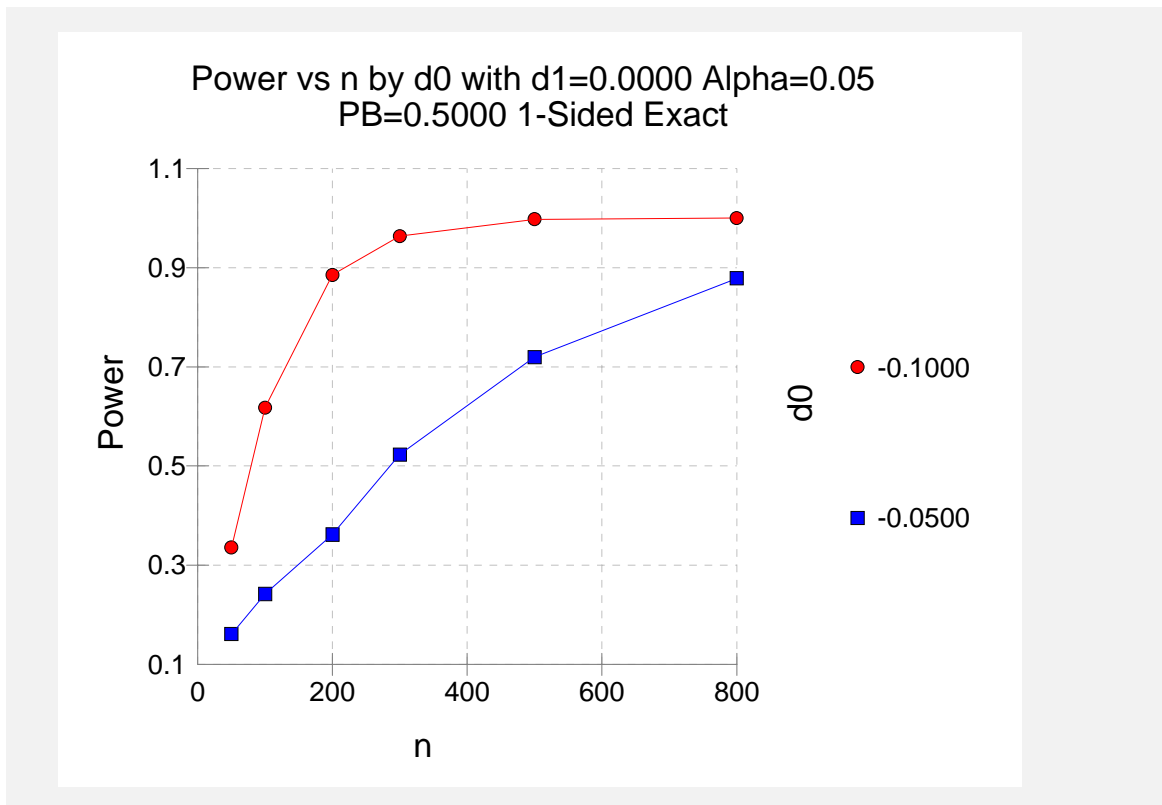
## Beta

Beta is the probability of accepting a false null hypothesis. It is the opposite of power.

## Reject H0 if $R \geq \text{This}$

The symbol  $R$  stands for the number of items with the characteristic of interest out of the  $n$  items sampled. For the scenario presented on the first line, an exact test does not exist for these parameters at the target alpha of 0.05. The closest that can be achieved is an alpha of 0.0314. In this case, we would reject the null hypothesis in any sample of size 50 in which the count of individuals with the characteristic of interest is greater than or equal to 27.

## Plots Section



This plot shows the relationship between power, sample size, and the trivial difference. Note that 90% power is achieved with an  $n$  of about 200 when the trivial difference is  $-.10$  and about 800 when the trivial difference is  $-.05$ .

## Example 2 - Finding the Sample Size

Continuing from Example 1, suppose you want to find the exact sample size necessary to achieve 90% power when the equivalence difference is -.05. Assume that an exact binomial test will be used.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. You can make these changes directly on your screen or you can load the Example2 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>n</b>
Higher Proportions Are .....	<b>Better</b>
Equivalence Difference (d0).....	<b>-.05</b>
Actual Difference (d1) .....	<b>0</b>
Baseline Proportion (PB) .....	<b>0.50</b>
n (Sample Size) .....	<i>Ignored since this is the Find setting</i>
N (Population Size) .....	<b>Infinite</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.10</b>
Test Statistic in Reports .....	<b>Exact Test</b>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>
Show Comparative Reports .....	<b>Not checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots.....	<b>Not checked</b>

### Annotated Output

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

#### Numeric Results

Numeric Results when H0:  $P = P_0$  versus H1:  $P = P > P_0$   
Test Statistic: Exact Test

Power	N	Equiv. Difference (d0)	Actual Difference (d1)	Baseline Proportion (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 If R>=This
0.9024	861	-0.0500	0.0000	0.5000	0.0500	0.0499	0.0976	412

This report shows that a sample size of 861 will be necessary to achieve the design requirements.

## Example 3 – Comparing Test Statistics

Continuing Example 1, suppose the researchers want to investigate which of the five test statistics to use. This is an important question since choosing the wrong test statistic may increase the sample size, reduce power or inflate the actual alpha level. The differences in the characteristics of test statistics are most noticeable in small samples. Hence, the investigation done here is for sample sizes of 20 to 200 in steps of 20. The trivial difference will be set to -.10. All other settings are as given in Example 1.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Higher Proportions Are.....	<b>Better</b>
Equivalence Difference (d0) .....	<b>-.10</b>
Actual Difference (d1) .....	<b>0</b>
Baseline Proportion (PB).....	<b>0.50</b>
n (Sample Size).....	<b>20 to 200 by 20</b>
N (Population Size) .....	<b>Infinite</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Not checked</b>
Show Comparative Reports .....	<b>Checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots.....	<b>Checked</b>

### Annotated Output

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

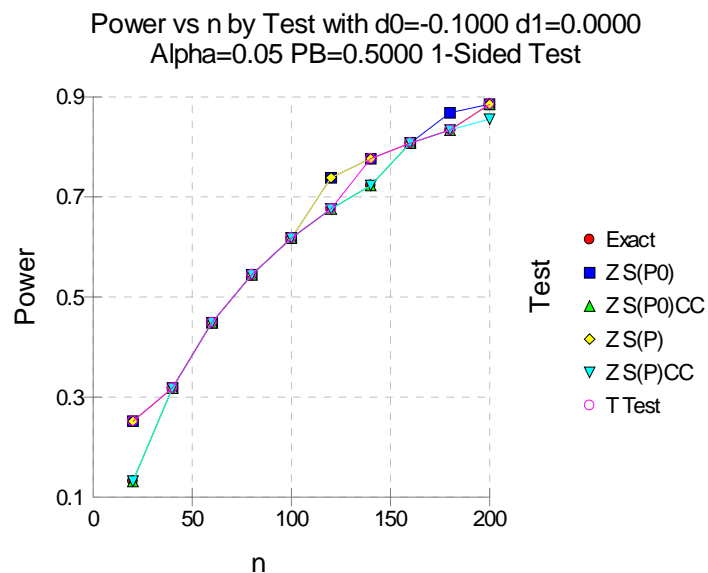
## Numeric Results

Power Comparison for Methods of Testing  $H_0: P = P_0$  versus  $H_1: P > P_0$

N	Equiv. Diff. (d0)	Actual Diff. (d1)	Baseline Prop. (PB)	Target Alpha	Exact Test Power	Z-Test S(P0) Power	Z-Test S(P0)C Power	Z-Test S(P) Power	Z-Test S(P)C Power	T-Test Power
20	-0.1000	0.0000	0.5000	0.0500	0.1316	0.2517	0.1316	0.2517	0.1316	0.2517
40	-0.1000	0.0000	0.5000	0.0500	0.3179	0.3179	0.3179	0.3179	0.3179	0.3179
60	-0.1000	0.0000	0.5000	0.0500	0.4487	0.4487	0.4487	0.4487	0.4487	0.4487
80	-0.1000	0.0000	0.5000	0.0500	0.5445	0.5445	0.5445	0.5445	0.5445	0.5445
100	-0.1000	0.0000	0.5000	0.0500	0.6178	0.6178	0.6178	0.6178	0.6178	0.6178
120	-0.1000	0.0000	0.5000	0.0500	0.6759	0.7385	0.6759	0.7385	0.6759	0.6759
140	-0.1000	0.0000	0.5000	0.0500	0.7229	0.7765	0.7229	0.7765	0.7229	0.7765
160	-0.1000	0.0000	0.5000	0.0500	0.8077	0.8077	0.8077	0.8077	0.8077	0.8077
180	-0.1000	0.0000	0.5000	0.0500	0.8337	0.8683	0.8337	0.8337	0.8337	0.8337
200	-0.1000	0.0000	0.5000	0.0500	0.8854	0.8854	0.8854	0.8854	0.8556	0.8854

Actual Alpha Comparison for Methods of Testing  $H_0: P = P_0$  versus  $H_1: P > P_0$

N	Equiv. Diff. (d0)	Actual Diff. (d1)	Baseline Prop. (PB)	Target Alpha	Exact Test Alpha	Z-Test S(P0) Alpha	Z-Test S(P0)C Alpha	Z-Test S(P) Alpha	Z-Test S(P)C Alpha	T-Test Alpha
20	-0.1000	0.0000	0.5000	0.0500	0.0210	0.0565	0.0210	0.0565	0.0210	0.0565
40	-0.1000	0.0000	0.5000	0.0500	0.0392	0.0392	0.0392	0.0392	0.0392	0.0392
60	-0.1000	0.0000	0.5000	0.0500	0.0445	0.0445	0.0445	0.0445	0.0445	0.0445
80	-0.1000	0.0000	0.5000	0.0500	0.0445	0.0445	0.0445	0.0445	0.0445	0.0445
100	-0.1000	0.0000	0.5000	0.0500	0.0423	0.0423	0.0423	0.0423	0.0423	0.0423
120	-0.1000	0.0000	0.5000	0.0500	0.0392	0.0575	0.0392	0.0575	0.0392	0.0392
140	-0.1000	0.0000	0.5000	0.0500	0.0358	0.0514	0.0358	0.0514	0.0358	0.0514
160	-0.1000	0.0000	0.5000	0.0500	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459
180	-0.1000	0.0000	0.5000	0.0500	0.0408	0.0558	0.0408	0.0558	0.0408	0.0408
200	-0.1000	0.0000	0.5000	0.0500	0.0492	0.0492	0.0492	0.0492	0.0363	0.0492



The first numeric report shows the power for each test statistic. The second shows the actual alpha achieved by the design.

An examination of the first report shows that the power is often different for at least one of the tests. Also notice that the exact test always has the minimum power in each row. This would lead us to discard this test statistic. However, consider the second report which shows the actual alpha level (the target was 0.05) for each test. By inspecting corresponding entries in both tables, we



see that whenever a test statistic achieves a better power than the exact test, it also yields an actual alpha level larger than the target alpha.

For example, look at the powers for  $n = 20$ . The z test using  $s(p \text{ hat})$  has an unusually large power = 0.2517. This is a much larger power than the exact test's value of 0.1316. However, note that the actual alpha for this test is 0.0560 which is larger than the target alpha of 0.05.

We conclude that indeed, the exact test is consistently the best test since it always achieves a significance level that is less than the target value.

## Example 4 - Validation using Chow, Shao, and Wang

The only appropriate example we have found is Chow, Shao, and Wang (2003) page 85, which gives the result of a sample size calculation using an asymptotic formula. They calculate a sample size of 22 when  $\alpha = 0.05$ ,  $\beta = 0.20$ ,  $PB = 0.30$ , Equiv. difference =  $-0.10$ , and actual proportion = 0.50. As we shall see, **PASS** obtains a different answer.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. You can make these changes directly on your screen or you can load the Example4 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>n</b>
Higher Proportions Are.....	<b>Better</b>
Equivalence Difference (d0) .....	<b>-0.10</b>
Actual Difference (d1) .....	<b>0.20</b>
Baseline Proportion (PB).....	<b>0.30</b>
n (Sample Size).....	<i>Ignored since this is the Find setting</i>
N (Population Size) .....	<b>Infinite</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.20</b>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>

### Annotated Output

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

## Numeric Results

Power	N	Equiv. Difference (d0)	Actual Difference (d1)	Baseline Proportion (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 If R>=This
0.8338	17	-0.1000	0.5000	0.3000	0.0500	0.0377	0.1662	7

**PASS** calculated a sample size of only 17 while Chow's formula obtained 22. The difference occurs because **PASS** uses exact calculations based on the binomial distribution, while Chow et al. use a large-sample approximation based on the normal approximation to the binomial. To see that 17 is indeed the correct answer, enter the values into **PASS**'s one-sample proportion test. The necessary values are  $P0 = 0.20$ ,  $P1 = 0.50$ ,  $\alpha = 0.05$ , and  $\beta = 0.10$ . These values result in a sample size of 17.

We have found that the approximate results are closer to the exact results when the sample sizes are over 200. For sample sizes less than 50, there can be significant error in the approximate formulas.

## Example 5 - Finding Power after an Experiment

The proportion of successes of the current treatment is known to be 0.74 based on years of treatment use. Researchers have developed a new method of treatment which costs about half the current treatment price. Before the new treatment can be approved it must be shown that the success of the proposed treatment is not inferior to that of the current treatment. It is determined that use of the new treatment is justifiable if it is shown that it is effective more than 70% of the time. Sixty individuals are randomly selected to receive the new method of treatment. Forty-three (71.67%) of the 60 individuals responded positively to the treatment. The p-value for the test based on exact binomial probabilities is 0.4514. Because the researchers were unable to show the new treatment is non-inferior, they desire to know the power of the test.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Higher Proportions Are.....	<b>Better</b>
Equivalence Difference (d0) .....	<b>-.04</b>
Actual Difference (d1) .....	<b>0.00</b>
Baseline Proportion (PB).....	<b>0.74</b>
n (Sample Size).....	<b>60</b>
N (Population Size) .....	<b>Infinite</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored</i>
Test Statistic in Reports .....	<b>Exact Test</b>

### Annotated Output

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

### Numeric Results

Numeric Results for testing  $H_0: P = P_0$  versus  $H_1: P > P_0$   
Test Statistic: Exact Test

Power	N	Equiv. Difference (d0)	Actual Difference (d1)	Baseline Proportion (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 If R>=This
0.1112	60	-0.0400	0.0000	0.7400	0.0500	0.0295	0.8888	49

This report shows that the power for the test was only 0.1112.

## Example 6 - Finding Sample Size Based on the Odds Ratio

## Chapter 110

# One Proportion – Equivalence

This module provides power analysis and sample size calculation for equivalence tests in one-sample designs in which the outcome is binary. Users may choose from among commonly-used test statistics.

The details of sample size calculations for the one-sample design for binary outcomes are presented in the chapter entitled One Proportion - Equality and will not be repeated here. Instead, this chapter discusses those changes necessary for equivalence tests.

Approximate sample size formulas for equivalence tests of a single proportion are presented in Chow et al. (2003) page 83. However, only large sample (normal approximation) results are given there. The results available in this module use exact calculations based on the enumeration of all possible values for the binomial distribution.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Example

An equivalence test example will set the stage for the discussion of the terminology that follows. Suppose that the current treatment for a disease is effective 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the questions that must be answered is whether the new treatment is equivalent to the current treatment. In other words, do about 70% of treated subjects respond to the new treatment?

It is known that the new treatment will not have a response rate that is exactly the same as that of the standard treatment. After careful consideration, they decide that the margin of equivalence is plus or minus 10%. That is, if the response rate of the new treatment is between 60% and 80% it will be deemed equivalent to the standard treatment.

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is within 10% of the standard (baseline) treatment. The statistical hypotheses to be tested are

$$H_0: |P - PB| \geq 0.1 \text{ versus } H_1: |P - PB| < 0.1$$

Notice that when the null hypothesis is rejected the conclusion is that the response rate is between 0.6 and 0.8.

## Technical Details

In the discussion that follows, let  $P$  represent the proportion being investigated. That is,  $P$  is the actual probability of a *success* in a binomial experiment. Often, this proportion is a response rate, cure rate, or survival rate. Let  $PB$  represent the *baseline* proportion. In an equivalence trial, the baseline proportion is the response rate of the current (standard) treatment. Let  $PL$  represent the smallest value of  $P$  that still results in the conclusion that the new treatment is equivalent to the current treatment. Similarly, let  $PU$  represent the largest value of  $P$  that still results in the conclusion that the new treatment is equivalent to the current treatment. Note that  $PB$  will be between  $PL$  and  $PU$ . The power of a test is computed at a specific value of the proportion,  $P_1$ .

The statistical hypotheses that are tested are

$$H0: P \leq PL \text{ or } P \geq PU \text{ versus } H1: PL < P < PU$$

This unusual hypothesis test can be broken down into two, one-sided hypothesis tests (TOST) as follows

$$H0: P \leq PL \text{ versus } H1: P > PL$$

and

$$H0: P \geq PU \text{ versus } H1: P < PU$$

If both of these one-sided tests are rejected at significance level  $\alpha$ , then equivalence can be concluded at significance level  $\alpha$ . Note that we do not conduct the individual tests at  $\alpha / 2$ .

There are three common methods of specifying the margin of equivalence. The most direct is to simply assign values for PL and PU. However, it is often more meaningful to identify PB and then specify PL and PU implicitly by giving a difference, ratio, or odds ratio. Mathematically, the definitions of these parameterizations are

<b><u>Parameter</u></b>	<b><u>Computation</u></b>	<b><u>Hypotheses</u></b>
Difference	$d0 =  PL - P_B  = PU - PB$	$H0: d \geq d0 \text{ vs } H1: d < d0$
Ratio	$r0 = \frac{1}{PL / PB} = PU / PB$	$H0: r \geq r0 \text{ vs } H1: r < r0$
Odds Ratio	$o0 = \frac{1}{OddsL / OddsB} = \frac{OddsU}{OddsB}$	$H0: o \geq o0 \text{ vs } H1: o < o0$

where

Difference	$d =  P - PB $
Ratio	$r = \begin{cases} P / PB & \text{if } P > PB \\ PB / P & \text{if } P < PB \end{cases}$
Odds Ratio	$o = \begin{cases} Odds / OddsB & \text{if } P > PB \\ OddsB / Odds & \text{if } P < PB \end{cases}$

## Difference

The difference is perhaps the most direct method of comparison between two proportions. It is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its use.

One difficulty arises when the event of interest is rare. If a difference of 0.001 occurs when the baseline probability is 0.40, it would be dismissed as being trivial. That is, there is usually little interest in a treatment that only decreases the probability from 0.400 to 0.399. However, if the baseline probability of a disease is 0.002, a 0.001 decrease would represent a reduction of 50%. Thus, interpretation of the difference depends on the baseline probability of the event. As a rule of thumb, the difference is best suited for those cases in which.

## Equivalence Test using a Difference

The following example might be instructive. Suppose 60% of patients respond to the current treatment method ( $P_B = 0.60$ ). If the response rate of the new treatment is no less than five percentage points different ( $d_0 = 0.05$ ) from the existing treatment, it will be considered to be equivalent. Substituting these figures into the statistical hypotheses gives

$$H_0: d \geq 0.05 \text{ versus } H_1: d < 0.05$$

where  $d = |P - P_B|$ .

The resulting joint hypotheses are

$$H_0: P \leq 0.55 \text{ versus } H_1: P > 0.55.$$

and

$$H_0: P \geq 0.65 \text{ versus } H_1: P < 0.65.$$

In this example, when both null hypotheses are rejected, the concluded alternative is that the response rate is between 55% and 65%.



## Ratio

The ratio  $r_0 = P_E / P_B$  denotes the relative change in the probability of the response. Testing equivalence uses the hypotheses

$$H_0: r \leq r_0 \text{ versus } H_1: r > r_0$$

where  $r = P / P_B$  if  $P > P_B$  or  $r = P_B / P$  if  $P < P_B$ .

## Equivalence Test using a Ratio

The following example might help to understand the concept of *equivalence* as defined by the ratio. Suppose that 60% of patients ( $P_B = 0.60$ ) respond to the current treatment method. If a new treatment changes the response rate by no more than 10% ( $r_0 = 1.1$ ), it will be considered to be equivalent to the standard treatment. Substituting these figures into the statistical hypotheses gives

$$H_0: r \geq 1.1 \text{ versus } H_1: r < 1.1$$

The relationship  $P_0 = (r_0)(P_B)$  gives the two, one-sided, hypotheses

$$H_0: P \leq 0.54 \text{ versus } H_1: P > 0.54$$

$$H_0: P \geq 0.66 \text{ versus } H_1: P < 0.66$$

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is between 54% and 66%.

## Odds Ratio

The odds ratio,  $o_0 = (PE / (1 - PE)) / (PB / (1 - PB))$ , gives the relative change in the odds of the response. Testing noninferiority and superiority use the same formulation, namely

$$H_0: o \leq o_0 \text{ versus } H_1: o > o_0$$

where  $o = \text{Odds} / \text{Odds}_B$  if  $P > P_B$  or  $o = \text{Odds}_B / \text{Odds}$  if  $P < P_B$ .

## Power and Sample Size Calculation

Historically, power and sample size calculations for a one-sample proportion test have been based on normal approximations to the binomial. However, with the speed of modern computers, using the normal approximation is unnecessary, especially for small samples. Rather, the significance level and power can be computed using complete enumeration of all possible values of  $x$ , the number of successes in a sample of size  $n$ .

This is done as follows.

1. The critical value of the test is computed using standard techniques.
2. For each possible value of  $x$ , the value of the test statistic ( $z$  test,  $t$  test, or exact test) is computed along with its associated probability of occurrence.
3. The significance level and power are computed by summing the probabilities of occurrence for all values of the test statistic that are greater than (or less than) the critical value. Each probability of occurrence is calculated using  $P_0$  for the significance level and  $P_1$  for the power.

Other variables such as the sample size are then found using an efficient search algorithm. Although this method is not as elegant as a closed-form solution, it is completely accurate.

## Examples of Power Calculation for the Exact Test

Suppose the baseline proportion,  $P_B$ , is 0.50, the sample size is 10, and the target alpha level is 0.05. A typical value for the equivalence difference is 0.05. However, because the example is for a small sample size, the equivalence difference will be set to 0.4 (which is, of course, a very unrealistic figure) for illustrative purposes. Calculate the power of this design to detect equivalence if the actual difference between the proportions is 0.10.

The first step is to find the rejection region under the null hypothesis. In this example, the null hypothesis is  $H_0: P \leq 0.1$  or  $H_0: P > 0.9$  and the alternative hypothesis is  $H_1: 0.1 < P < 0.9$ . This composite hypothesis breaks down into the following two, one-sided, simple hypotheses

1.  $H_0: P \leq 0.1$  versus  $H_1: P > 0.1$
2.  $H_0: P \geq 0.9$  versus  $H_1: P < 0.9$

The rejection regions for the both tests are determined from the following table of cumulative binomial probabilities for  $N = 10$ . The first column of probabilities is for  $r$  greater than or equal to  $R$  while the second two columns of probabilities are for  $r$  less than or equal to  $R$ .

**Table of Binomial Probabilities for  $N = 10$  and  $P = 0.1, 0.9$ , and  $0.6$**

R	Pr( $r \geq R$   $P=0.1$ )	Reject Test1	Pr( $r \leq R$   $P=0.9$ )	Reject Test2	Reject Both	Pr( $r \leq R$   $P=0.6$ )
0	1.0000	No	0.0000	Yes	No	0.0001
1	0.6513	No	0.0000	Yes	No	0.0017
2	0.2639	No	0.0000	Yes	No	0.0123
3	0.0702	No	0.0000	Yes	No	<b>0.0548</b>
4	<b>0.0128</b>	Yes	0.0001	Yes	Yes	0.1662
5	0.0016	Yes	0.0016	Yes	Yes	0.3669
6	0.0001	Yes	<b>0.0128</b>	Yes	Yes	<b>0.6177</b>
7	0.0000	Yes	0.0702	No	No	0.8327
8	0.0000	Yes	0.2639	No	No	0.9536
9	0.0000	Yes	0.6513	No	No	0.9940
10	0.0000	Yes	1.0000	No	No	1.0000

The second column gives the value of alpha for the first test ( $H_0: P \leq 0.1$  versus  $H_1: P > 0.1$ ). The rejection region for this test is all values of  $R$  greater than or equal to 4. The fourth column gives the values of alpha for the second test. The rejection region for the second test is all values of  $R$  less than or equal to 6. The rejection region for both tests is those values of  $R$  values that result in rejection of both individual tests. These are the  $R$  values 4, 5, and 6. The power is computed using the final column of the table which gives cumulative binomial probabilities for  $P = 0.5 + 0.1 = 0.6$ . The power is probability for the cases 4, 5, and 6. It is calculated as  $0.6177 - 0.0548 = 0.5629$ .

It is informative to consider what happens when the equivalence difference is reduced from 0.4 to 0.2. The following table gives the appropriate cumulative binomial probabilities for this case.

**Table of Binomial Probabilities for  $N = 10$  and  $P = 0.3, 0.7$ , and  $0.6$**

R	Pr( $r \geq R$   $P=0.3$ )	Reject Test1	Pr( $r \leq R$   $P=0.7$ )	Reject Test2	Reject Both	Pr( $r \leq R$   $P=0.6$ )
0	1.0000	No	0.0000	Yes	No	0.0001
1	0.9718	No	0.0001	Yes	No	0.0017
2	0.8507	No	0.0016	Yes	No	0.0123
3	0.6172	No	0.0106	Yes	No	0.0548
4	0.3504	No	<b>0.0473</b>	Yes	No	0.1662
5	0.1503	No	0.1503	No	No	0.3669
6	<b>0.0473</b>	Yes	0.3504	No	No	0.6177
7	0.0106	Yes	0.6172	No	No	0.8327
8	0.0016	Yes	0.8507	No	No	0.9536
9	0.0001	Yes	0.9718	No	No	0.9940
10	0.0000	Yes	1.0000	No	No	1.0000

The second column gives the value of alpha for the first test. The rejection region for this test is all values of  $R$  greater than or equal to 6. The fourth column gives the values of alpha for the second test. The rejection region for the second test is all values of  $R$  less than or equal to 4. The rejection region for both tests together is empty! There is no  $R$  for which both tests will be rejected. Hence, the alpha level and the power will both be 0.0.

## Examples of Power Calculation for the Z Test

The following example illustrates how to calculate the power of an approximate  $z$  test. There are several  $z$  tests to choose from. We will use the following test.

$$z = \frac{p - P_0}{\sqrt{P_0(1 - P_0) / n}}$$

Calculating the rejection region for the  $z$  test is based on a table of normal probabilities. For the target alpha level of 0.05, the critical value is 1.6449. That is, the first hypothesis test that  $H_0: P \leq 0.1$  versus  $H_1: P > 0.1$  is rejected if the resulting calculated  $z$  value is greater than 1.6449. Similarly, the second hypothesis test that  $H_0: P \geq 0.9$  versus  $H_1: P < 0.9$  is rejected when the calculated  $z$  value is less than -1.6449. The rejection regions for the both tests are shown in the following table of binomial probabilities for  $N = 10$ .

**Table Showing Both One-Sided Z Tests for  $N = 10$  and  $P = 0.1, 0.9$ , and  $0.6$**

R	Z for P = 0.1	Reject Test1	Z for P = 0.9	Reject Test2	Reject Both	Pr( $r \leq R   P = 0.6$ )
0	-1.0541	No	-9.4868	Yes	No	0.0001
1	0.0000	No	-8.4327	Yes	No	0.0017
2	<b>1.0541</b>	No	-7.3786	Yes	No	<b>0.0123</b>
3	2.1082	Yes	-6.3246	Yes	Yes	0.0548
4	3.1623	Yes	-5.2705	Yes	Yes	0.1662
5	4.2164	Yes	-4.2164	Yes	Yes	0.3669
6	5.2705	Yes	-3.1623	Yes	Yes	0.6177
7	6.3246	Yes	-2.1082	Yes	Yes	<b>0.8327</b>
8	7.3786	Yes	-1.0541	No	No	0.9536
9	8.4327	Yes	0.0000	No	No	0.9940
10	9.4868	Yes	1.0541	No	No	1.0000

Note that the null hypothesis is rejected for the equivalence test when  $R$  is 3, 4, 5, 6, and 7. The power is the probability of these values calculated using  $P = 0.60$ . It is calculated as  $0.8327 - 0.0123 = 0.8204$ . Notice that this is much larger than 0.5629 which was the power for the exact test. The reason for this discrepancy is that the approximate test is actually testing at a larger alpha than the target of 0.05. The actual alpha is the maximum of the two individual alphas. From the first table, we can see that the actual alpha for the first test is  $\Pr(r \geq 3 | P = 0.1) = 0.0702$ . Similarly, the actual alpha for the second test is  $\Pr(r \leq 7 | P = 0.9) = 0.0702$ . Hence the alpha level is 0.0702. The actual alpha of the exact test was 0.0128.

## Test Statistics

The test statistics used are given in the chapter entitled One Proportion - Equality. They will not be repeated 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Alpha*, *Beta*, and *n*. In most situations, you will select either Beta or *n*.

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.

### Baseline Proportion (PB)

Enter a value (or range of values) for the baseline proportion. In an equivalence study, this is the response rate of the standard (existing) treatment. Note that this is not the value of  $P_0$ . Instead, this value is used in the calculation of  $P_0$ .

Proportions must be between zero and one.

You may enter a range of values such as *0.1, 0.2, 0.3* or *0.1 to 0.9 by 0.1*.

### n (Sample Size)

Enter a value (or range of values) for the sample size *n*. This is the number of individuals sampled in the study. Values must be integers greater than one.

You may enter a range such as *10, 50, 100* or *10 to 100 by 10*.

### Alpha (Significance Level)

This option specifies one or more values for alpha which is the probability of a type-I (false positive) error. This error occurs when you falsely reject the null hypothesis.

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. You should pick a value for alpha that represents the risk of a type-I you are willing to take in your experimental situation.

Note that because of the discrete nature of the binomial distribution, the alpha level will not usually be achieved exactly.

### Beta (1 - Power)

This option specifies one or more values for beta which is the probability of a type-II (false negative) error. This 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, but now 0.1 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. 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.

## Test Statistic in Report

Specify which test statistic will be used in searching and reporting. Note that *C.C.* is an abbreviation for *Continuity Correction*. This refers to the adding or subtracting of  $1/(2n)$  to (or from) the numerator of the z-value to bring the normal approximation closer to the binomial distribution.

In most situations, you would select the 'Exact Test' option. The other options are provided for comparative purposes.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is given directly in terms of the proportions PL, PU, PB, and P1.

### Actual Proportion (PB)

This is the value of the proportion, P1 at which the power is calculated. The power calculations assume that this is the actual value of the proportion. For noninferiority tests, this value is often set equal to PB.

Proportions must be between zero and one. You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

### Upper and Lower Equivalence Proportion

These options set the smallest and largest values which are still to be considered trivially different from PB. Note that the lower proportion must be less than PB, and the upper proportion must be greater than PB. Since these values are proportions, they must be positive values less than one. They cannot be equal to PB.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference of two proportions.

### Equivalence Difference (d0)

This option sets the smallest value which is still trivially different from PB by setting the magnitude of the difference between P0 and PB. For example, if PB (baseline proportion) is 0.50, you might consider differences of 0.01, 0.02, or 0.04 to be small enough so that the fact that P0 is different from 0.50 can be overlooked. However, you might decide that if the difference is 0.05 or more, the treatment is not equivalent. Thus, this value would be set to 0.05.

Since this value is an absolute difference between two proportions, it must be between 0 and 1.

### Actual Difference (d1)

This option specifies the value of P1 (the actual proportion) by specifying the difference between the two proportions, P1 and PB. This difference is used with PB to calculate the value of P1 using the formula:  $P1 = PB + \text{difference}$ . For equivalence tests, this value is often set equal to zero.

Differences must be between -1 and 1.

You may enter a range of values such as *.03 .05 .10* or *.01 to .05 by .01*.

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio of two proportions.

### Equivalence Ratio (r0)

This option sets the value which is still trivially different from PB by setting the ratio between P0 and PB. For P0 example, if PB (baseline proportion) is 0.50, you might consider ratios of 0.99, 0.98, or even 0.96 to be small enough so that the fact that P0 is less than PB can be overlooked (the difference is trivial). However, you might decide that if the ratio is 0.95 or less, the treatment is not equivalent. Thus, this value would be set to 0.95.

Since this value is a ratio between two proportions, it must be positive. Since it is a margin, it cannot be one. Also, it cannot be so large that the calculated value of P0 is greater than one.

### Actual Ratio (r1)

This option specifies the value of P1 (the actual proportion) by specifying the ratio between the two proportions, P1 and PB. This ratio is used with PB to calculate the value of P1 using the formula:  $P1 = (\text{Ratio})(PB)$ . For equivalence tests, this value is often set equal to one.

Ratios must be greater than zero. Note that the ratios must be small enough so that P1 is less than one.

You may enter a range of values such as *.5 .6 .7 .8* or *1.25 to 2.0 by .25*.



## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratios.

### Equivalence Odds Ratio (o0)

This option sets the value which is still trivially different from PB by setting the odds ratio of P0 and PB. For example, if PB (baseline proportion) is 0.50, you might consider odds ratios of 0.99, 0.98, or even 0.96 to be small enough so that the fact that P0 is less than PB can be overlooked (the difference is trivial). However, you might decide that if the odds ratio is 0.80 or less, the treatment is inferior. Thus, this value would be set to 0.80.

Since this value is a ratio between two odds, it must be positive. Because it is a margin, it cannot be one.

### Actual Odds Ratio (o1)

This option specifies the value of P1 (the actual proportion) by specifying the odds ratio between the two proportions, P1 and PB. This ratio is used with PB to calculate the value of P1. For noninferiority tests, this value is often set equal to one.

Odds ratios must be greater than zero. You may enter a range of values such as *.5 .6 .7 .8* or *1.25 to 2.0 by .25*.

## Options Tab

The Options tab allows for specification of the maximum number of iterations to be used in searches.

### Maximum Iterations during Searches

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported.

## Example1 - Finding the Power

Suppose 50% of patients with a certain type of cancer survive two years using the current treatment. The current treatment is expensive and has several severe side effects. A new treatment has fewer side effects and is less expensive. An equivalence trial is to be conducted to show that the two-year survival rate of the new treatment is the same as the current treatment. After serious consideration, the margin of equivalence is set at 5%. What power will be achieved by sample sizes of 50, 100, 200, 300, 500, or 800 and a significance level of 0.05? For comparative purposes, also calculate the power for margin of equivalence of 10%. Assume that the true survival rate of the new treatment is the same as that of the current (baseline) treatment.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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 .....	<b>Beta and Power</b>
Baseline Proportion (PB) .....	<b>0.50</b>
n (Sample Size) .....	<b>50 100 200 300 500 800</b>
Actual Difference (d1) .....	<b>0</b>
Equivalence Difference (d0).....	<b>0.05 0.10</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
Test Statistic in Reports .....	<b>Exact Test</b>

##### Reports Tab

Show Numeric Reports .....	<b>Checked</b>
Show Comparative Reports .....	<b>Not checked</b>
Show Definitions .....	<b>Checked</b>
Show Plots .....	<b>Checked</b>
Show Comparative Plots.....	<b>Not checked</b>

### Annotated Output

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

## Numeric Results

**Numeric Results when H0: Non-Equivalence versus H1: Equivalence**  
**Test Statistic: Exact Test**

Power	N	Equiv. Diff. (d0)	Lower Equiv. Prop. (P0L)	Upper Equiv. Prop. (P0U)	Actual Diff. (d1)	Baseline Prop. (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 if R1<=R<=R2 (R1 R2)
0.0000	50	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0000	1.0000	29 21
0.0000	100	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0000	1.0000	54 46
0.0000	200	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0000	1.0000	103 97
0.0460	300	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0465	0.9540	150 150
0.4390	500	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0484	0.5610	244 256
0.7567	800	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0476	0.2433	384 416
0.0000	50	0.1000	0.4000	0.6000	0.0000	0.5000	0.0500	0.0000	1.0000	27 23
0.2356	100	0.1000	0.4000	0.6000	0.0000	0.5000	0.0500	0.0423	0.7644	49 51
0.7708	200	0.1000	0.4000	0.6000	0.0000	0.5000	0.0500	0.0492	0.2292	92 108
0.9267	300	0.1000	0.4000	0.6000	0.0000	0.5000	0.0500	0.0443	0.0733	135 165
0.9952	500	0.1000	0.4000	0.6000	0.0000	0.5000	0.0500	0.0461	0.0048	219 281
0.9999	800	0.1000	0.4000	0.6000	0.0000	0.5000	0.0500	0.0453	0.0001	344 456

### Report Definitions

Power is the probability of concluding equivalence when the proportions are equivalent.

N is the size of the sample drawn from the population.

The equivalence difference is the maximum value of the difference that is still considered unimportant.

The actual difference is the value of the difference under the alternative hypothesis.

PB is the baseline or standard value of the proportion. This is the value under the current treatment.

P0L and P0U are the limits between which an equivalent proportion must fall.

d0 is the smallest absolute difference that is still considered equivalent.

d1 is the value of the difference under the alternative hypothesis.

Alpha is the probability of concluding equivalence when the proportions are non-equivalent.

Beta is the probability concluding non-equivalence when the proportions are equivalent.

### Summary Statements

A sample size of 50 achieves 0% power to detect a difference (P0-PB) of 0.0500 using a two-sided binomial test. The target significance level is 0.0500. The actual significance level achieved by this test is 0.0000. These results assume a baseline proportion (PB) of 0.5000 and that the actual difference (P1-PB) is 0.0000.

This report shows the values of each of the parameters, one scenario per row. Because of the discrete nature of the binomial distribution, the target alpha is usually different than the actual alpha. Hence, the actual alpha is also shown.

## Power

Power is the probability of concluding equivalence when the treatment is indeed equivalent.

## N

This is the sample size.

## Equivalence Difference (or Proportion, Ratio, or Odds Ratio)

The equivalence difference is the maximum difference from the baseline proportion, PB, that is still considered as unimportant or trivial. This value is used to calculate P0.

## Equivalence Upper and Lower Proportions

If the true proportion is between these two limits, the treatment is considered to be equivalent to the baseline proportion. These are the bounds of equivalence.

## Actual Difference (or Proportion, Ratio, or Odds Ratio)

The actual difference is the difference between the actual proportion,  $P_1$ , and the baseline proportion,  $P_B$ .

## Baseline Proportion

The baseline proportion,  $P_B$ , is the response rate that is achieved by the current (standard) treatment.

## Target Alpha

This is the target (set in the design) value of the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. That is, this is the probability of concluding equivalence when in fact the new treatment is not equivalent. Because of the discreteness of the binomial distribution from which this value is calculated, the target value is seldom achieved.

## Actual Alpha

This is the actual value of alpha (see Target Alpha) that is achieved by the design. Note that low values of alpha reduce the power.

## Beta

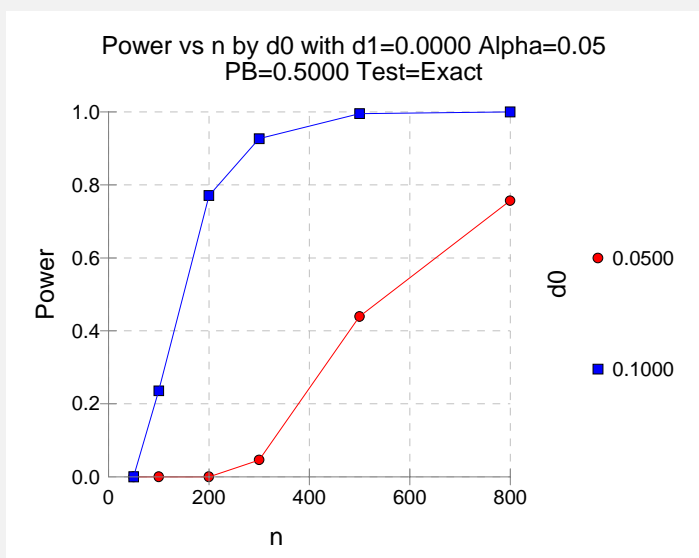
Beta is the probability of accepting a false null hypothesis. It is the opposite of power.

## Reject $H_0$ if $R_1 \leq R \leq R_2$

This value provides the bounds between which equivalence is concluded. For example, if  $n$  is 50, then a value here of 29|31 means that the null hypothesis of non-equivalence is rejected when the number of items with the characteristic of interest is 29, 30, or 31.

When the second number is less than the first as it is in the first line (29|21), the design can never reject the null hypothesis. These designs should never be used.

## Plots Section



This plot shows the relationship between power, sample size, and the trivial difference. Note that 80% power is achieved with a sample size of about 210 when the trivial difference is 0.10 and over 800 when the trivial difference is 0.05.

## Example2 - Finding the Sample Size

Continuing from Example1, suppose you want to find the exact sample size necessary to achieve 90% power when the trivial difference is 0.05. Assume that an exact binomial test will be used.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>n</b>
Baseline Proportion (PB).....	<b>0.50</b>
Actual Difference (d1) .....	<b>0</b>
Equivalence Difference (d0) .....	<b>0.05</b>
n .....	<i>Ignored since this is the Find setting</i>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
Test Statistic in Reports .....	<b>Exact Test</b>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>
Show Comparative Reports .....	<b>Not checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots .....	<b>Not checked</b>

### Annotated Output

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

#### Numeric Results

**Numeric Results when H0: Non-Equivalence versus H1: Equivalence**  
Test Statistic: Exact Test

Power	N	Equiv. Diff. (d0)	Lower Equiv. Prop. (P0L)	Upper Equiv. Prop. (P0U)	Actual Diff. (d1)	Baseline Prop. (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 if R1<=R<=R2 (R1 R2)
0.9040	1092	0.0500	0.4500	0.5500	0.0000	0.5000	0.0500	0.0498	0.0960	519 573

This report shows that a sample size of 1092 will be necessary to achieve the design requirements.

## Example3 – Comparing Test Statistics

Continuing Example1, suppose the researchers want to investigate which of the five test statistics to use. This is an important question since choosing the wrong test statistic can increase sample size and reduce power. The differences in the characteristics of test statistics are most noticeable in small samples. Hence, the investigation done here is for sample sizes of 20 to 200 in steps of 20. The trivial difference will be set to 0.10. All other settings are as given in Example 1.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Baseline Proportion (PB) .....	<b>0.50</b>
Actual Difference (d1) .....	<b>0</b>
Equivalence Difference (d0).....	<b>0.10</b>
n .....	<b>20 to 200 by 20</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Not checked</b>
Show Comparative Reports .....	<b>Checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots.....	<b>Checked</b>

### Annotated Output

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



The first report shows the power for each test statistic. The second report shows the actual alpha achieved by the design.

An examination of the first report shows that once non-zero powers are obtained, they are often different for at least one of the tests. Also notice that the exact test always has the minimum power in each row. This would lead us to discard this test statistic. However, consider the second report which shows the actual alpha level (the target was 0.05) for each test. By inspecting corresponding entries in both tables, we see that whenever a test statistic achieves a better power than the exact test, it also yields an actual alpha level larger than the target alpha.

For example, look at the powers for  $n = 120$ . The  $z$  test using  $s(P_0)$  has an unusually large power = 0.4770. This is a much larger power than the exact test's value of 0.3517. However, note that the actual alpha for this test is 0.0575 which is larger than the target alpha of 0.05 and the exact test's alpha of 0.0392.

We conclude that indeed, the exact test is consistently the best test since it always achieves a significance level that is less than the target value.



## Example4 - Validation

We could not find a worked example for this situation in the literature. Therefore, we will use the example that was worked 'by hand' earlier in this chapter to validate the program. In that example, the baseline proportion was 0.50, alpha was 0.05,  $n$  was 10, the actual difference was 0.10, and the trivial difference was 0.40. The power was calculated to be 0.5629.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. You can make these changes directly on your screen or you can load the Example4 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Baseline Proportion (PB).....	<b>0.50</b>
Actual Difference (d1) .....	<b>0.10</b>
Equivalence Difference (d0) .....	<b>0.40</b>
n .....	<b>10</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>

# Annotated Output

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

## Numeric Results

Numeric Results when H0: Non-Equivalence versus H1: Equivalence										
Test Statistic: Exact Test										
Power	N	Equiv. Diff. (d0)	Lower Equiv. Prop. (P0L)	Upper Equiv. Prop. (P0U)	Actual Diff. (d1)	Baseline Prop. (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 if R1<=R<=R2 (R1 R2)
0.5630	10	0.4000	0.1000	0.9000	0.1000	0.5000	0.0500	0.0128	0.4370	4 6

PASS has obtained the same answer within rounding error.

## Example5 - Computing the Power after Completing the Experiment

Researchers are testing a generic drug to determine if it is equivalent to the name-brand alternative. Equivalence is declared if the success rate of the generic brand is no more than 10% from that of the name-brand drug. Suppose that the name-brand drug is known to have a success rate of 60%. In a study of 500 individuals, they find that 265, or 53%, are successfully treated using the generic brand. An equivalence test (exact test) with  $\alpha = 0.05$  failed to declare that the two drugs are equivalent. The researchers would now like to compute the power for actual differences ranging from 0 to 9%.

Note that the power is not calculated solely at the difference observed in the study, 7%. It is more informative to study a range of values with practical significance.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Baseline Proportion (PB).....	<b>0.60</b>
Actual Difference (d1) .....	<b>0.0 to 0.09 by 0.01</b>
Equivalence Difference (d0) .....	<b>0.10</b>
n .....	<b>500</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
Test Statistic in Reports .....	<b>Exact Test</b>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>
Show Comparative Reports .....	<b>Not checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Checked</b>
Show Comparative Plots .....	<b>Not checked</b>

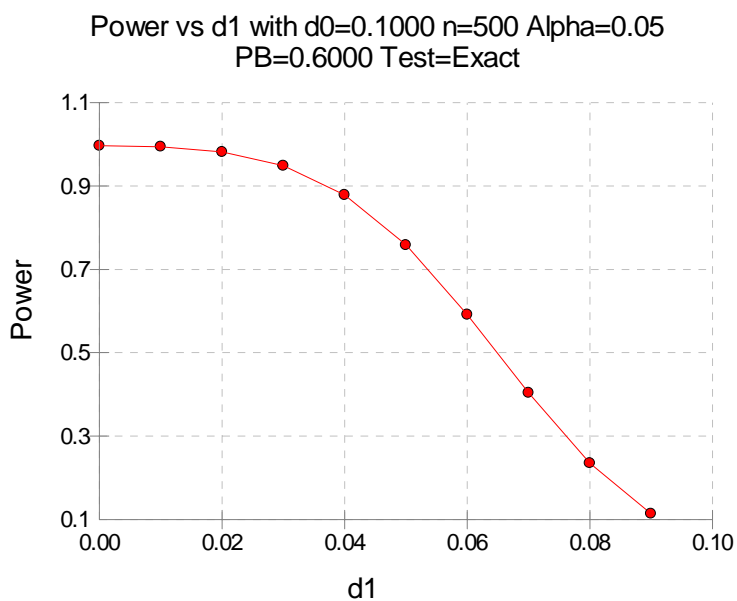
### Annotated Output

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

## Numeric Results

Numeric Results when H0: Non-Equivalence versus H1: Equivalence  
Test Statistic: Exact Test

Power	N	Equiv. Diff. (d0)	Lower Equiv. Prop. (P0L)	Upper Equiv. Prop. (P0U)	Actual Diff. (d1)	Baseline Prop. (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 if R1<=R<=R2 (R1 R2)
0.9965	500	0.1000	0.5000	0.7000	0.0000	0.6000	0.0500	0.0489	0.0035	269 332
0.9940	500	0.1000	0.5000	0.7000	0.0100	0.6000	0.0500	0.0489	0.0060	269 332
0.9815	500	0.1000	0.5000	0.7000	0.0200	0.6000	0.0500	0.0489	0.0185	269 332
0.9482	500	0.1000	0.5000	0.7000	0.0300	0.6000	0.0500	0.0489	0.0518	269 332
0.8783	500	0.1000	0.5000	0.7000	0.0400	0.6000	0.0500	0.0489	0.1217	269 332
0.7583	500	0.1000	0.5000	0.7000	0.0500	0.6000	0.0500	0.0489	0.2417	269 332
0.5914	500	0.1000	0.5000	0.7000	0.0600	0.6000	0.0500	0.0489	0.4086	269 332
0.4041	500	0.1000	0.5000	0.7000	0.0700	0.6000	0.0500	0.0489	0.5959	269 332
0.2352	500	0.1000	0.5000	0.7000	0.0800	0.6000	0.0500	0.0489	0.7648	269 332
0.1139	500	0.1000	0.5000	0.7000	0.0900	0.6000	0.0500	0.0489	0.8861	269 332



The range in power is quite large. The power is relatively high and constant if the true difference is less than or equal to 4%, but it decreases rapidly as the differences increase from there.

## Example6 - Finding the Sample Size using Ratios

Researchers would like to compare a new treatment to an existing standard treatment. The new treatment will be deemed equivalent to the standard treatment if the response rate is changed by no more than 20%, hence,  $r = 1.20$ . It is known that 60% of patients respond to the standard treatment. If the researchers use the exact test and a significance level of 0.05, how large of a sample must they take to achieve 90% power if the actual ratio is 1.0?

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. You can make these changes directly on your screen or you can load the Example6 template by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>n</b>
Baseline Proportion (PB).....	<b>0.60</b>
Actual Ratio (r1) .....	<b>1.0</b>
Equivalence Ratio (r0).....	<b>1.2</b>
n .....	<i>Ignored since this is the Find setting</i>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
Test Statistic in Reports .....	<b>Exact Test</b>
<b>Reports Tab</b>	
Show Numeric Reports .....	<b>Checked</b>
Show Comparative Reports .....	<b>Not checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Checked</b>
Show Comparative Plots .....	<b>Not checked</b>

# Annotated Output

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

## Numeric Results

**Numeric Results when H0: Non-Equivalence versus H1: Equivalence**  
Test Statistic: Exact Test

Power	N	Equiv. Ratio (r0)	Lower Equiv. Prop. (P0L)	Upper Equiv. Prop. (P0U)	Actual Ratio (r1)	Baseline Prop. (PB)	Target Alpha	Actual Alpha	Beta	Reject H0 if R1<=R<=R2 (R1 R2)
0.9014	228	1.2000	0.5000	0.7200	1.0000	0.6000	0.0500	0.0488	0.0986	127 152

**Summary Statements**  
A sample size of 228 achieves 90% power to detect a ratio (P0/PB) of 1.2000 using a two-sided binomial test. The target significance level is 0.0500. The actual significance level achieved by this test is 0.0488. These results assume a baseline proportion (PB) of 0.6000 and that the actual ratio (P1/PB) is 1.0000.

They must sample 228 individuals to achieve just over 90% power for an actual ratio of 1.0 and trivial ratio of 1.20.

## Chapter 115

# Confidence Interval for a Proportion

## Introduction

This routine calculates the sample size necessary to achieve a required precision at a stated confidence coefficient for a confidence interval about the sample proportion.

## Technical Details

Rather than use the normal approximation to the binomial distribution, *PASS* uses exact calculations based on the binomial distribution as presented in Desu and Raghavarao (1990). The results will be similar to the normal approximation results when the sample size is large (greater than 200). However, when the sample size is small, the normal approximation is poor. The point to remember is that *PASS* uses the exact calculations, not approximations as some other programs do.

Let  $X$  be a discrete random variable with probability function

$$f(x; P0) = P0^x (1 - P0)^{1-x}$$

where  $x = 0, 1$ .  $X$  is said to follow the Bernoulli distribution. Let  $X1, X2, \dots, Xn$  be a random sample of  $X$  of size  $n$ . Then

$$Y = \sum_{i=1}^n X_i$$

has a binomial distribution with parameters  $n$  and  $P0$ . The probability distribution of  $Y$  is

$$f(y; n, P0) = \binom{n}{y} P0^y (1 - P0)^{1-y}$$

We note that  $Y/n$  is the usual estimate of  $P0$ .

To estimate  $P0$  while controlling the absolute error with a given probability, you find  $n$  to satisfy

$$\Pr\left(\left|\frac{Y}{n} - P0\right| < d\right) \geq 1 - \alpha$$

Here,  $d$  represents the precision of the estimate. This probability statement may be rearranged as follows:

$$\begin{aligned}\Pr\left(\left|\frac{Y}{n} - P0\right| < d\right) &= \Pr\left(\left[n(P0 - d)\right] + 1 \leq Y \leq \left[n(P0 + d)\right]\right) \\ &= \Pr(y_1 \leq Y \leq y_2) \\ &= \sum_{y=y_1}^{y_2} \binom{n}{y} P0^y (1 - P0)^{n-y}\end{aligned}$$

Hence, for given values of  $d$ ,  $P0$ , and  $1 - \alpha$ , a search can be made for the smallest value of  $n$  for which the above probability statement is true. This is what is done in this routine.

## Finite Population Size

The above calculations assume that samples are being drawn from a large (infinite) population. When the population is of finite size ( $N$ ), a normal approximation to the hypergeometric distribution is used. Calculations are based on the formula

$$z_{1-\alpha/2} = \frac{d}{\sqrt{\frac{P0(1-P0)(N-n)}{nN}}}$$

## 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 proportion 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 ( $d$  in the formulas). That is, the confidence interval is formed by taking the sample proportion plus and minus this amount. This is sometimes called the *margin of error*. The smaller this amount, the more precise will be the interval.

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

### 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  $\alpha$ . 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.

## P0 (Baseline Proportion)

This is an estimate of the population proportion. Since the point of the study is to estimate this value, you will not know it in advance. However, since it is necessary to estimate this value, you must put in something. The largest sample sizes occur when  $P0 = 0.50$ . So when you do not have a good estimate of  $P0$ , use a value near 0.5. Your sample size will then be a little larger than necessary.

Remember that  $P0 + d$  must be less than one and that  $P0 - d$  must be greater than zero ( $d$  is the precision).

## Options Tab

This tab sets a couple of options 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.

### Iterative Precision

When a search is made for the precision value, this is the cutoff value used to terminate the search. In most cases, a value of 0.0001 will be more than sufficient.

# Example1 - Calculating Sample Size

Suppose a study is planned in which you want to construct a confidence interval for the proportion that is no wider than 0.02. You want a confidence coefficient of 0.95. Previous studies show that the proportion is about 0.4. Instead of looking at a precision of just 0.02 with a confidence coefficient of 0.95, you want to see the sample sizes for a range of values from 0.02 to 0.07 and at two confidence coefficients, 0.95 and 0.99. Calculate the necessary sample size.

## Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N (Sample Size)</b>
Precision .....	<b>0.02 to 0.07 by 0.01</b>
Confidence Coefficient .....	<b>0.95, 0.99</b>
Population Size .....	<b>Infinite</b>
N (Sample Size) .....	<b>Ignored</b>
P0 (Baseline Proportion) .....	<b>0.4</b>

## Annotated Output

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

### Numeric Results

Numeric Results			
	C.C.	N	P0
Precision	Confidence Coefficient	Sample Size	Baseline Proportion
0.02000	0.95073	2281	0.40000
0.02000	0.99014	3955	0.40000
0.03000	0.95051	1005	0.40000
0.03000	0.99023	1754	0.40000
0.04000	0.95165	566	0.40000
0.04000	0.99003	980	0.40000
0.05000	0.95208	365	0.40000
0.05000	0.99056	634	0.40000
0.06000	0.95023	244	0.40000
0.06000	0.99058	435	0.40000
0.07000	0.95175	181	0.40000
0.07000	0.99045	315	0.40000

#### Report Definitions

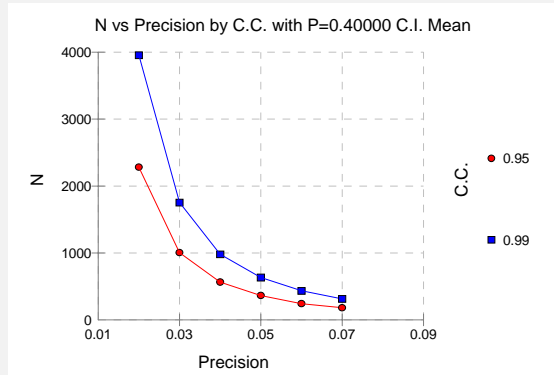
Precision is the plus and minus value used to create the confidence interval.  
 Confidence Coefficient is probability value associated with the confidence interval.  
 N is the size of the sample drawn from the population.  
 P0 is the estimated baseline proportion.

#### Summary Statements

A sample size of 2281 produces a 95% confidence interval equal to the sample proportion plus or minus 0.02000 when the estimated proportion is 0.40000.

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 Hahn and Meeker

Hahn and Meeker (1991) page 145 state that a sample size of 30 gives a 90% confidence interval of 0.50 plus or minus 0.16. We will run this problem through *PASS*.

## Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Confidence Coefficient</b>
Precision .....	<b>0.02 to 0.07 by 0.01</b>
Confidence Coefficient.....	<b>0.95, 0.99</b>
Population Size .....	<b>Infinite</b>
N (Sample Size) .....	<b>30</b>
P0 (Baseline Proportion) .....	<b>0.5</b>

## 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	P0 Baseline Proportion
0.16000	0.90126	30	0.50000

The confidence coefficient calculated by *PASS* is also 0.90.

## Chapter 120

# Single-Stage Phase II Clinical Trials

## Introduction

Phase II clinical trials determine whether a drug or regimen has sufficient activity against disease to warrant more extensive study and development. In a single-stage design, a single group of patients is studied. Usually, investigators will know the response rate of other drugs against the disease. Unless the current drug can be shown to be significantly more effective, its use will not be pursued.

This module finds designs that meet the error rate ( $\alpha$  and  $\beta$ ) criterion and minimize the sample size when an exact test of proportions is used. The algorithm, discussed by A'Hern (2001), is an exact version of the algorithm of Fleming (1982)

## Technical Details

*Phase I clinical trials* are designed to provide information about the maximum tolerated dose levels of a treatment. They consist of three to six patients at each dose level and provide little information about the effectiveness of the treatment.

*Phase II trials* obtain initial estimates of the degree of treatment activity. A patient's response may be measured by the decrease in the size of a tumor. For example, a patient may be considered to have responded to treatment if the tumor shrinks by 50% or more. There is no control group in these designs. Rather, the purpose of the trial is to determine if the drug shows enough activity against disease to warrant a full-scale, phase III clinical trial.

Let  $P_0$  be the largest response proportion that, if true, clearly implies that the treatment does not warrant further study.  $P_0$  is sometimes called the response rate of a poor treatment. For example, for a new anti-tumor drug, this may be set to 0.10.

Let  $P_1$  be the smallest response proportion that, if true, clearly implies that the treatment does warrant further study.  $P_1$  is sometimes called the response rate of a good treatment. For example, for a new anti-tumor drug, this may be set to 0.30.

A statistical test of hypothesis may be conducted to test the null hypothesis that  $P \leq P_0$  versus the alternative hypothesis that  $P \geq P_1$  ( $P$  is the true proportion responding to the treatment in the population). Let  $\alpha$  be the probability of rejecting the null hypothesis when it is true. Let  $\beta$  be the probability of rejecting the alternative hypothesis when it is true.

A single-stage phase II design can be represented by two numbers:  $N$  and  $R$ .  $N$  is the sample size.  $R$  is the critical value. If  $R$  or fewer responses occur in the  $N$  patients, the drug is rejected. The design is found by searching for the minimum value of  $N$  for which a value for  $R$  can be found such that the following two error rate constraints are met:

$$\Pr(\text{reject} | P_0, R, N) \geq 1 - \alpha$$

and

$$\Pr(\text{reject} | P_1, R, N) \leq \beta$$

### Limiting the Range of the Search

Because of the discrete nature of the binomial distribution by which these error rates are calculated, there is no closed-form solution and so a search among possible values of  $N$  must be conducted. In order to speed up the search, only values of  $N$  between  $0.8F$  and  $4F$  are considered.  $F$  is the sample size based on the normal approximation to the binomial, suggested by Fleming (1982).

# 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.

## Data Tab

### P0 (Maximum Response of a Poor Treatment)

Enter one or more response proportions of a poor drug. If the true proportion responding to the treatment is less than this amount, study of the treatment will not be recommended.

This value must be less than  $P1$  and greater than zero.

### P1 (Minimum Response of a Good Treatment)

Enter one or more response proportions of a good drug. If the true proportion responding to the treatment is greater than or equal to this amount, study of the treatment can be recommended.

This value must be greater than  $P0$  and less than one.

## Alpha

Alpha is the probability of rejecting the hypothesis that the proportion responding to the treatment is less than or equal to  $P0$  when this hypothesis is actually true. That is,  $\text{Alpha} = \Pr(\text{Rejecting } P \leq P0 \mid P \leq P0)$ .

The range of Alpha is 0.001 to 0.25. Popular values are 0.05 and 0.10.

## Beta

This is the probability of rejecting the hypothesis that the proportion responding to the treatment is greater than or equal to  $P1$  when this hypothesis is true. That is,  $\text{Beta} = \Pr(\text{rejecting } P \geq P1 \mid P \geq P1)$ .

The range of Beta is 0.001 to 0.4. Popular values are 0.10 and 0.20.

## Example1 - Validation from A'Hern

A'Hern (2001) presents tables of sample sizes for various values of the design parameters. Setting  $\alpha = 0.05$ ,  $\beta = 0.20$ ,  $P_0 = 0.05$ , and  $P_1 = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ , and  $0.9$ , A'Hern finds the corresponding sampling plans to be (using the notation  $R+1/N$ ) 14/169, 4/27, 3/14, 2/7, 2/5, 2/4, 2/4, 1/1, and 1/1. This would be set up as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
P0.....	<b>0.05</b>
P1.....	<b>.1 .2 .3 .4 .5 .6 .7 .8 .9</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

#### Single Stage Design for Testing $H_0: P \leq P_0$ versus $H_1: P \geq P_1$

<b>P0</b>	<b>P1</b>	<b>Alpha</b>	<b>Beta</b>	<b>Cut-Off R + 1</b>	<b>N</b>	<b>Actual Alpha</b>	<b>Actual Beta</b>
0.050	0.100	0.050	0.200	14	169	0.045	0.194
0.050	0.200	0.050	0.200	4	27	0.044	0.182
0.050	0.300	0.050	0.200	3	14	0.030	0.161
0.050	0.400	0.050	0.200	2	7	0.044	0.159
0.050	0.500	0.050	0.200	2	5	0.023	0.188
0.050	0.600	0.050	0.200	2	4	0.014	0.179
0.050	0.700	0.050	0.200	2	4	0.014	0.084
0.050	0.800	0.050	0.200	1	1	0.050	0.200
0.050	0.900	0.050	0.200	1	1	0.050	0.100

#### Report Definitions

P0 is the maximum response proportion of a poor drug.

P1 is the minimum response proportion of a good drug.

N is the sample size.

If the number of responses  $\geq R+1$ , P0 is rejected.

If the number of responses  $\leq R$ , P1 is rejected.

Alpha is the probability of rejecting that  $P \leq P_0$  when this is true.

Beta is the probability of rejecting that  $P \geq P_1$  when this is true.

#### Summary Statements

A study requires 169 subjects to decide whether the proportion responding, P, is less than or equal to 0.050 or greater than or equal to 0.100. If the number of responses is 14 or more, the hypothesis that  $P \leq 0.050$  is rejected with a target error rate of 0.050 and an actual error rate of 0.045. If the number of responses is 13 or less, the hypothesis that  $P \geq 0.100$  is rejected with a target error rate of 0.200 and an actual error rate of 0.194.

Note that the designs match those of A'Hern (2001) exactly.



## Chapter 125

# Two-Stage Phase II Clinical Trials

## Introduction

Phase II clinical trials determine whether a drug or regimen has sufficient activity against disease to warrant more extensive study and development. In a two-stage design, the patients are divided into two groups or stages. At the completion of the first stage, an interim analysis is made to determine if the second stage should be conducted. If the number of patients responding is greater than a certain amount, the second stage is conducted. Otherwise, it is not.

This module finds designs that meet the error rate (alpha and beta) criterion and minimize the expected sample size. The algorithm is discussed in Simon (1989). Extending Simon's work, our algorithm allows the investigation of near-optimal designs that may have other useful properties.

## Technical Details

*Phase I clinical trials* are designed to provide information about the maximum tolerated dose levels of a treatment. They consist of three to six patients at each dose level and provide little information about the effectiveness of the treatment.

*Phase II trials* obtain initial estimates of the degree of treatment activity. A patient's response may be measured by the decrease in the size of a tumor. For example, a patient may be considered to have responded to treatment if the tumor shrinks by 50% or more. There is no control group in these designs. Rather, the purpose of the trial is to determine if the drug shows enough activity against disease to warrant a full-scale, phase III clinical trial.

Let  $P_0$  be the largest response proportion which, if true, clearly implies that the treatment does not warrant further study.  $P_0$  is sometimes called the response rate of a poor treatment. For a new anti-tumor drug, this may be set to 0.10.

Let  $P_1$  be the smallest response proportion which, if true, clearly implies that the treatment does warrant further study.  $P_1$  is sometimes called the response rate of a good treatment. For a new anti-tumor drug, this may be set to 0.30.

A statistical test of hypothesis may be conducted to test the null hypothesis that  $P \leq P_0$  versus the alternative hypothesis that  $P \geq P_1$  ( $P$  is the true proportion responding to the treatment in the population). Let  $\alpha$  be the probability of rejecting the null hypothesis when it is true. Let  $\beta$  be the probability of rejecting the alternative hypothesis when it is true.

A phase II design can be represented by four numbers:  $NI$ ,  $RI$ ,  $N$ , and  $R$ .  $NI$  is the sample size in the first stage.  $RI$  is the critical value in the first stage. If  $RI$  or fewer responses occur in the  $NI$  patients, the drug is rejected.  $N$  is the combined sample size for both the first and second stages.  $R$  is the critical value in the combined sample. If  $R$  or fewer of the  $N$  patients respond, the drug is rejected.

The expected (or average) sample size of this design is

$$E(N_E) = N1 + (1 - PET)(N - N1)$$

where  $PET$  is the probability of early termination of the study.

The probability of rejecting a drug with success probability  $P$  can be found using the binomial distribution. The formulation is

$$\Pr(\text{reject} | P, N1, R1, R, N) = B(R1 | P, N1) + \sum_{X=R1+1}^{\min(N1, R)} b(X | P, N1) B(R - X | P, N - N1)$$

where

$$b(X | P, N) = \frac{N!}{X!(N - X)!} P^X (1 - P)^{N - X}$$

and

$$B(X | P, N) = \sum_{r=0}^X b(r | P, N)$$

The two error rate constraints are

$$\Pr(\text{reject} | P0, N1, R1, R, N) \geq 1 - \alpha$$

and

$$\Pr(\text{reject} | P1, N1, R1, R, N) \leq \beta$$

### Optimum Design

The optimum design minimizes the average sample size,  $E(N)$ , while meeting the error rate constraints. This design is found through an exhaustive search of all possible designs. This search may take several minutes to complete.

### Designs Other Than Optimal

The optimal design minimizes the average sample size. There are examples where a less-than optimal design may be more desirable. For example, suppose the optimal design were  $N1 = 5$  and  $N = 25$ . This design is poor because only 5 patients are obtained during the first stage, but 20 are needed during the second stage. Most researchers would rather have more balance in the sample sizes of the two stages. Because of this, the actual optimal design may be rejected on other grounds.

## Design Flexibility

Dealing with sequential designs is complicated. It may be difficult to achieve exactly the number of patients proscribed for each phase. However, it should be remembered that the validity of the probability statements depends on the sample size requirements being met exactly. This is because the interpretation of an error rate probability statement is for repeated studies conducted in exactly the same way. We envision that if many studies of the same drug are conducted using the specific sampling plan  $NI, RI, N, R$  when  $P = P0$ , a proportion  $\alpha$  of them will be falsely terminated due to chance occurrences.

The point is, the interpretation of the error rates is for a large number of identical studies in which the sampling plan is identical and as proscribed. If the sampling plan is allowed to vary, this interpretation is invalid. Of course, the degree of possible error in interpretation depends on the degree to which the sampling plan is changed. We recognize that when dealing with human subjects, flexibility must be maintained. However, the scientist must also recognize that when the sampling plan is changed, the exact probability statements can no longer be calculated.

## 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.

## Which Designs are Displayed

This parameter specifies which designs are displayed. Since several thousand designs may be considered during the search for the optimum, it is important to limit the number of designs reported on.

The options are:

### All designs

All designs considered are output. This option should only be used in special cases in which a small number of designs are tested. Otherwise, hundreds of pages of output will be generated.

### Only designs that meet alpha & beta constraints

Only designs that meet the alpha and beta constraints are shown. This allows you to consider many near optimal designs which may be selected on grounds other than expected sample size.

### Optimum designs only

Only the optimum design, the minimax design, and the single stage design are displayed.

## P0 (Poor)

This is the response proportion of a poor drug. If the true proportion responding to the treatment is less than this amount, study of the treatment would not be recommended.

This value must be less than  $P1$  and greater than zero.

Only one value can be entered.

## P1 (Good)

This is the response proportion of a good drug. If the true proportion responding to the treatment is greater than or equal to this amount, study of the treatment would be recommended.

This value must be greater than  $P0$  and less than one.

Only one value can be entered.

## Alpha

Alpha is the probability of rejecting the hypothesis that the proportion responding to the treatment is less than or equal to  $P0$  when this hypothesis is actually true. That is,  $\text{Alpha} = \text{Pr}(\text{Rejecting } P \leq P0 | P \leq P0)$ .

The range of Alpha is 0.001 to 0.25. Popular values are 0.05 and 0.10.

## Beta

This is the probability of rejecting the hypothesis that the proportion responding to the treatment is greater than or equal to  $P1$  when this hypothesis is true. That is,  $\text{Beta} = \text{Pr}(\text{rejecting } P \geq P1 | P \geq P1)$ .

The range of Beta is 0.001 to 0.4. Popular values are 0.10 and 0.20.

## N Min

$N$  is the combined sample size of the two stages of the design. This parameter sets the minimum value of  $N$  that is used during the search. The optimum value of  $N$  must be between  $N$  Min and  $N$  Max or it will not be found.

The keyword MIN indicates that the value used is the minimum of the smallest sample size from a single stage design and  $MIN2$  where  $MIN2$  is calculated using

$$MIN2 = \frac{P_0 + P_1}{2} \left( 1 - \frac{P_0 + P_1}{2} \right) \left[ \frac{z_{1-\alpha} + z_{1-\beta}}{P_1 - P_0} \right]^2$$

Since it is unlikely that the two stage sample size will be less MIN, this provides a reasonable starting point for a search for  $N$ . **However, experience has shown that you should use a small number such as 2 to insure that you obtain the optimum.**

You can also enter a value like MIN- $x$  where  $x$  is a positive integer. This will cause the search to begin  $x$  units below the MIN.

The problem here is that this procedure may take a long time to run. Specifying a good starting value significantly reduces the running time.

Examples of valid entries are 2, 10, 20, MIN, MIN-1, MIN-15.

## N Max

$N$  is the combined sample size of the two stages of the design. This parameter sets the maximum value of  $N$  used during the search. The optimum value of  $N$  should be between  $N$  Min and  $N$  Max or it will not be found.

The keyword BEST+ $X$  indicates that the search should try at least  $X$  units above the latest optimum value of  $N$ . For example, suppose the  $N$  Min is set at 10. The search algorithm begins at 10, and then continues by examining 11, 12, and so. Suppose that the search finds a candidate optimum at  $N = 13$ . To make sure that 13 is the optimum, the search continues on from 13 to  $13+X$  (if, for example,  $X = 5$ , this value is 18). If no new optimum designs are found, the design at  $N = 13$  is selected.

When using this option,  $X$  should be set large enough to guarantee that the true optimum can be found, but small enough so that the search does not take hours to complete. Our experience is that  $X$  should be greater than or equal to 8.

Examples of valid entries are for this parameter are:

20  
30  
BEST+8  
BEST 8 (the plus sign is optional)  
BEST 3  
Best 4 (capitalization is not necessary)

## N Step

This parameter sets the step size in the search for  $N$ . Usually, you would enter 1 here. Occasionally, you may want to increase the step size during the initial part of your search to speed up convergence. Once you have determined a likely range, you can tighten up the search boundaries and reset this value to 1.

## R Min

$R$  is the treatment rejection number for the combined samples. If the total number of patients responding to the treatment is not greater than  $R$ , the treatment is deemed unworthy of further study.  $R$  Min sets the lower boundary for  $R$  during the search for the optimum design. The optimum design must have an  $R$  value between  $R$  Min and  $R$  Max.

The recommended value for this parameter is zero. Its range is from zero to  $N$ .

## R Max

$R$  is the rejection number for the combined samples. If the total number of patients responding to the treatment is not greater than  $R$ , the treatment is deemed unworthy of further study.

$R$  Max sets the upper boundary for  $R$  during the search for the optimum design. The optimum design must have an  $R$  value between  $R$  Min and  $R$  Max.

Since the upper value is  $N$  and  $N$  is also a varying parameter, you can set this parameter to  $MAX$  or  $MAX-X$  (replacing  $X$  with an appropriate integer like 1, 2, or 3). This causes the maximum value of  $R$  to be set to the current value of  $N-X$  during each iteration of the search.

## R Step

This parameter sets the step size in the search for  $R$ . Usually, you would enter  $I$  here. Occasionally, you may want to increase the step size during the initial part of your search to speed up convergence. Once you have determined a likely range, you can tighten up the search boundaries and reset this value to 1.

## N1 Min

$N1$  is the sample size of the first stage. This value sets the minimum value of  $N1$  that is used during the search. The optimum value must be between  $N1$  Min and  $N1$  Max or it will not be found.

Although, in theory, the sample first stage design may have only 1, 2, or 3 patients, you may want to ignore such designs from consideration by setting this value to 4 or 5.

The actual range of this parameter is from 1 to  $N$ .

## N1 Max

$N1$  is the sample size of the first stage of the design. This parameter sets the maximum value of  $N1$  used during the search. The optimum value of  $N1$  should be between  $N$  Min and  $N$  Max or it will not be found. Although, in theory, the sample first stage design may have  $N-3$ ,  $N-2$ , or  $N-1$  patients, you may want to ignore such designs from consideration by setting this value to a smaller number.

Since the upper value is  $N-1$  and  $N$  is also a varying parameter, you can set this parameter to MAX or MAX- $X$  (replacing  $X$  with an appropriate value like 1 or 2). This causes the maximum value of  $N1$  to be set to the current value of  $N-X$ .

Examples:

10

20

MAX

Max-2

max-4

## N1 Step

This parameter sets the step size in the search for  $N1$ . Usually, you would enter a  $I$  here. Occasionally, you may want to increase the step size during the initial part of your search to speed up convergence. Once you have determined a likely range, you can tighten up the search boundaries and reset this value to 1.

## R1 Min

$R1$  is the drug rejection number for the first stage. If the number of patients responding to the treatment in the first stage is not greater than  $R1$ , the treatment is deemed unworthy of further study. This parameter sets the lower boundary for  $R1$  during the search for the optimum design. The optimum design must have an  $R1$  value between  $R$  Min and  $R$  Max.

The recommended value for this parameter is zero. Its range is from zero to  $N1$ .

## R1 Max

$R1$  is the rejection number for the first stage. If the number of patients responding to the treatment in the first stage is not greater than  $R1$ , the treatment is deemed unworthy of further study. This parameter sets the upper boundary for  $R1$  in the search for the optimum design.

Since the upper value is  $N1$  and  $N1$  is a varying parameter, you can set this parameter to  $MAX$  or  $MAX-X$  (replacing  $X$  with an integer like 1, 2, or 3). This causes the maximum value of  $R1$  to be set to the current value of  $N1-X$ .

The valid range of  $R1$  is between zero and  $N1$ .

## R1 Step

This parameter sets the step size in the search for  $R1$ . Usually, you would enter a 1 here.

Occasionally, you may want to increase the step size during the initial part of your search to speed up convergence. Once you have determined a likely range, you can tighten up the search boundaries and reset this value to 1.

## Options Tab

This tab sets a couple of options 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 the Power

Suppose a design is wanted for the case  $\text{Alpha} = 0.05$ ,  $\text{Beta} = 0.20$ ,  $P0 = 0.05$ , and  $P1 = 0.25$ . This would be set up as follows:

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Which Designs .....	<b>Optimum designs only</b>
P0 (Poor).....	<b>0.05</b>
P1 (Good) .....	<b>0.25</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>
N Min.....	<b>Min-1</b>
N Max.....	<b>Best+8</b>
N Step.....	<b>1</b>
R Min.....	<b>0</b>
R Max.....	<b>Max-3</b>
R Step.....	<b>1</b>
N1 Min.....	<b>1</b>
N1 Max.....	<b>Max-4</b>
N1 Step.....	<b>1</b>
R1 Min.....	<b>0</b>
R1 Max.....	<b>Max-1</b>
R1 Step.....	<b>1</b>



# Annotated Output

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

## Numeric Results

Possible Designs For  $P_0=0.050$ ,  $P_1=0.250$ ,  $\alpha=0.050$ ,  $\beta=0.200$

N1	R1	PET	N	R	Ave N	Alpha	Beta	Constraints Satisfied
16	2	0.000	16	2	16.00	0.043	0.197	Single Stage
12	0	0.540	16	2	13.84	0.043	0.199	Minimax
9	0	0.630	17	2	11.96	0.047	0.188	Optimum

### Report Definitions

N1 is the sample size in the first stage.

R1 is the drug rejection number in the first stage.

PET is the probability of early termination of the study.

N is the combined sample size of both stages.

R is the combined drug rejection number after both stages.

Ave N is the average sample size if this design is repeated many times.

Alpha is the probability of rejecting that  $P \leq P_0$  when this is true.

Beta is the probability of rejecting that  $P \geq P_1$  when this is true.

$P_0$  is the response proportion of a poor drug.

$P_1$  is the response proportion of a good drug.

### Summary Statements

The optimal two-stage design to test the null hypothesis that  $P \leq 0.050$  versus the alternative that  $P \geq 0.250$  has an expected sample size of 11.96 and a probability of early termination of 0.630. If the drug is actually not effective, there is a 0.047 probability of concluding that it is (the target for this value was 0.050). If the drug is actually effective, there is a 0.188 probability of concluding that it is not (the target for this value was 0.200). After testing the drug on 9 patients in the first stage, the trial will be terminated if 0 respond. If the trial goes on to the second stage, a total of 17 patients will be studied. If the total number responding is less than or equal to 2, the drug is rejected.

This report shows three designs. The first is the smallest single stage design. The second is the Minimax solution. This is the design with the smallest total sample size ( $N$ ). The third is the optimum design—the one that minimizes the average sample size.

## Example2 - Validation using Simon

Simon (1989) page 4 in his Table 1 presents designs for several scenarios. The first row of the table sets  $P0$  to 0.05,  $P1$  to 0.25, Alpha to 0.10, and Beta to 0.10. The optimal design is  $N1 = 9$ ,  $R1 = 0$ ,  $N = 24$ , and  $R = 2$ . The minimax design is  $N1 = 13$ ,  $R1 = 0$ ,  $N = 20$ , and  $R = 2$ . We will now run this example through *PASS*.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Which Designs .....	<b>Optimum designs only</b>
P0 (Poor) .....	<b>0.05</b>
P1 (Good) .....	<b>0.25</b>
Alpha .....	<b>0.10</b>
Beta .....	<b>0.10</b>
N Min .....	<b>Min-1</b>
N Max .....	<b>Best+8</b>
N Step .....	<b>1</b>
R Min .....	<b>0</b>
R Max .....	<b>Max-3</b>
R Step .....	<b>1</b>
N1 Min .....	<b>1</b>
N1 Max .....	<b>Max-4</b>
N1 Step .....	<b>1</b>
R1 Min .....	<b>0</b>
R1 Max .....	<b>Max-1</b>
R1 Step .....	<b>1</b>

### Annotated Output

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

#### Numeric Results

Possible Designs For P0=0.050, P1=0.250, Alpha=0.050, Beta=0.200								Constraints
N1	R1	PET	N	R	Ave N	Alpha	Beta	Satisfied
20	2	0.000	20	2	20.00	0.075	0.091	Single Stage
13	0	0.513	20	2	16.41	0.074	0.097	Minimax
9	0	0.630	24	2	14.55	0.093	0.097	Optimum

*PASS* has calculated exactly the same optimal design and minimax design.

## Chapter 130

# Three-Stage Phase II Clinical Trials

## Introduction

Phase II clinical trials determine whether a drug or regimen has sufficient activity against disease to warrant more extensive study and development. In a three-stage design, the patients are divided into three groups or stages. At the completion of the first stage, an interim analysis is made to determine if the second stage should be conducted. If the number of patients responding is greater than a certain amount, the second stage is conducted. Otherwise, it is not. A similar interim analysis is conducted at the end of the second stage.

This module finds designs that meet the error rate (alpha and beta) criterion and minimize the expected sample size. The formulation is given in Chen (1997). Extending Chen's work, our algorithm allows the investigation of near-optimal designs that may have other useful properties.

## Technical Details

*Phase I clinical trials* are designed to provide information about the maximum tolerated dose levels of a treatment. They consist of three to six patients at each dose level and provide little information about the effectiveness of the treatment.

*Phase II trials* obtain initial estimates of the degree of treatment activity. A patient's response may be measured by the decrease in the size of a tumor. For example, a patient may be considered to have responded to treatment if the tumor shrinks by 50% or more. There is no control group in these designs. Rather, the purpose of the trial is to determine if the drug shows enough activity against disease to warrant a full-scale, phase III clinical trial.

Let  $P_0$  be the largest response proportion which, if true, clearly implies that the treatment does not warrant further study.  $P_0$  is sometimes called the response rate of a *poor* treatment. For a new anti-tumor drug, this may be set to 0.10.

Let  $P_1$  be the smallest response proportion which, if true, clearly implies that the treatment does warrant further study.  $P_1$  is sometimes called the response rate of a *good* treatment. For a new anti-tumor drug, this may be set to 0.30.

A statistical test of hypothesis may be conducted to test the null hypothesis that  $P \leq P_0$  versus the alternative hypothesis that  $P \geq P_1$  ( $P$  is the true proportion responding to the treatment in the population). Let  $\alpha$  be the probability of rejecting the null hypothesis when it is true. Let  $\beta$  be the probability of rejecting the alternative hypothesis when it is true.

A three-stage phase II design can be represented by six numbers:  $R1$ ,  $N1$ ,  $R2$ ,  $N2$ ,  $R3$  and  $N3$ .  $N1$  is the sample size in the first stage.  $R1$  is the critical value in the first stage. If  $R1$  or fewer responses occur in the  $N1$  patients, the drug is rejected.  $N2$  is the total sample size of stages one and two.  $R2$  is the critical value in the second stage. If  $R2$  or fewer responses occur in the  $N2$  patients, the drug is rejected.  $N3$  is the combined sample size of all three stages.  $R3$  is the critical value in the combined sample. If  $R3$  or fewer of the  $N3$  patients respond, the drug is rejected.

The expected (or average) sample size of this design is

$$E(N_e) = N1 + (1 - PET1)(N2 - N1) + (1 - PET2)(N3 - N2)$$

where  $PET1$  is the probability of early termination of the study after stage one and  $PET2$  is the probability of early termination after stage two.

The probability of rejecting a drug with success proportion  $P$  can be found using the binomial distribution. The formulation is

$$\Pr(\text{reject} | P, N1, R1, R2, N2, R3, N3) = PET1 + PET2 + PET3$$

where

$$PET1 = B(R1 | P, N1)$$

$$PET2 = \sum_{X1=R1+1}^{\min(N1, R2)} b(X1 | P, N1) B(R2 - X1 | P, N2 - N1)$$

$$PET3 = \sum_{X1=R1+1}^{\min(N1, R3)} b(X1 | P, N1) \sum_{X2=R2+1-X1}^{\min(N3-N2, R3-X1)} b(X2 | P, N2 - N1) B(R3 - X1 - X2 | P, N3 - N2)$$

$$b(X | P, N) = \frac{N!}{X!(N-X)!} P^X (1-P)^{N-X}$$

$$B(X | P, N) = \sum_{r=0}^X b(r | P, N)$$

The two error rate constraints are

$$\Pr(\text{reject} | P0, N1, R1, R2, N2, R3, N3) \geq 1 - \alpha$$

and

$$\Pr(\text{reject} | P1, N1, R1, R2, N2, R3, N3) \leq \beta$$

### Optimum Design

The optimum design minimizes the average sample size,  $E(N)$ , while meeting the error rate constraints. This design is found through an exhaustive search of all possible designs. This search may take several minutes to complete.

## Designs Other Than Optimal

The optimal design minimizes the average sample size. There are examples where a less-than optimal design may be more desirable. For example, suppose the optimal design were  $N1 = 5$ ,  $N2 = 25$ , and  $N3 = 26$ . This design is poor because the bulk of the subjects are tested in the second phase. Most researchers would rather have more balance in the sample sizes of the three stages. For reasons like this, the actual optimal design may be replaced by another, sub-optimal, design.

## Design Flexibility

Dealing with sequential designs is complicated. It may be difficult to achieve exactly the number of patients proscribed for each phase. However, it should be remembered that the validity of the probability statements depends on the sample size requirements being met exactly. This is because the interpretation of an error rate probability statement is for repeated studies conducted in exactly the same way. We envision that if many studies of the same drug are conducted using the specific sampling plan when  $P = P0$ , a proportion  $\alpha$  of them will be falsely terminated due to chance occurrences.

The point is, the interpretation of the error rates is for a large number of identical studies in which the sampling plan is identical and as proscribed. If the sampling plan is allowed to vary, this interpretation is invalid. Of course, the degree of possible error in interpretation depends on the degree to which the sampling plan is changed. We recognize that when dealing with human subjects, flexibility must be maintained. However, the researcher must also recognize that when the sampling plan is changed, the exact probability statements can no longer be calculated.

# 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, 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.

## Which Designs are Displayed

This parameter controls which designs are displayed. Since several thousand designs may be considered during the search for the optimum, it is important to limit the number of designs reported on.

The options are:

### All designs

All designs considered are output. This option should only be used in special cases in which a small number of designs are tested. Otherwise, hundreds of pages of output will be generated.

### Only designs that meet alpha & beta constraints

Only designs that meet the alpha and beta constraints are shown. This allows you to consider many near optimal designs which may be selected on grounds other than expected sample size.

### Optimum designs only

Only the optimum design, the minimax design, and the single stage design are displayed.

## P0 (Poor)

This is the response proportion of a poor drug. If the true proportion responding to the treatment is less than this amount, study of the treatment would not be recommended.

This value must be less than  $P1$  and greater than zero.

Only one value can be entered.

## P1 (Good)

This is the response proportion of a good drug. If the true proportion responding to the treatment is greater than or equal to this amount, study of the treatment would be recommended.

This value must be greater than  $P0$  and less than one.

Only one value can be entered.

## Alpha

Alpha is the probability of rejecting the hypothesis that the proportion responding to the treatment is less than or equal to  $P0$  when this hypothesis is actually true. That is,  $\text{Alpha} = \text{Pr}(\text{Rejecting } P \leq P0 | P \leq P0)$ .

The range of Alpha is 0.001 to 0.25. Popular values are 0.05 and 0.10.

## Beta

This is the probability of rejecting the hypothesis that the proportion responding to the treatment is greater than or equal to  $P1$  when this hypothesis is true. That is,  $\text{Beta} = \text{Pr}(\text{rejecting } P \geq P1 | P \geq P1)$ .

The range of Beta is 0.001 to 0.4. Popular values are 0.10 and 0.20.

## N Min

$N$  is the combined sample size of the three stages of the design. This parameter sets the minimum value of  $N3$  that is used during the search. The optimum value of  $N3$  must be between  $N$  Min and  $N$  Max or it will not be found.

The keyword MIN indicates that the value used is the minimum of the smallest sample size from a single stage design and  $MIN2$  where  $MIN2$  is calculated using

$$MIN2 = \frac{p_0 + p_1}{2} \left( 1 - \frac{p_0 + p_1}{2} \right) \left[ \frac{z_{1-\alpha} + z_{1-\beta}}{p_1 - p_0} \right]^2$$

Since it is unlikely that the three stage sample size will be less MIN, this provides a reasonable starting point for a search for  $N$ . You can also enter a value like MIN- $x$  where  $x$  is a positive integer. This will cause the search to begin  $x$  units below the MIN.

This procedure may take a long time to run. Specifying a good starting value significantly reduces the running time.

Examples of valid entries are 2, 10, 20, MIN, MIN-1, MIN-15.

## N Max

$N$  is the combined sample size of the three stages of the design. This parameter sets the maximum value of  $N_3$  used during the search. The optimum value of  $N_3$  should be between N Min and N Max or it will not be found.

The keyword BEST+ $X$  indicates that the search should try at least  $X$  units above the latest optimum value of  $N_3$ . For example, suppose the N Min is set at 10. The search algorithm begins at 10, and then continues by examining 11, 12, and so. Suppose that the search finds a candidate optimum at  $N = 13$ . To make sure that 13 is the optimum, the search continues on from 13 to  $13+X$  (if, for example,  $X = 5$ , this value is 18). If no new optimum designs are found, the design at  $N_3 = 13$  is selected.

When using this option,  $X$  should be set large enough to guarantee that the true optimum can be found, but small enough so that the search does not take hours to complete. Our experience indicates that  $X$  should be greater than or equal to 8.

Examples of valid entries for this parameter are:

20  
30  
BEST+8  
BEST 8 (the plus sign is optional)  
BEST 3  
Best 4 (capitalization is not necessary)

## Options Tab

This tab sets a couple of options 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 results are left blank. A value of 500 is recommended.

# Example1 - Calculating the Power and Validation using Chen

Chen (1997) provides the minimax and optimum design for the case  $\alpha = 0.05$ ,  $\beta = 0.20$ ,  $P_0 = 0.05$ , and  $P_1 = 0.25$ . The optimum design is 0/8, 1/13, and 2/19. The minimax design is 0/12, 1/15, and 2/16.

## Setup

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

### Option Value

#### Data Tab

Which Designs ..... **Optimum designs only**

P0 (Poor) ..... **0.05**

P1 (Good) ..... **0.25**

Alpha ..... **0.05**

Beta ..... **0.20**

N Min ..... **Min**

N Max ..... **Best+2**

Note that the search may take several minutes to run, depending on the speed of your computer.

## Annotated Output

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

### Numeric Results

Possible Designs For  $P_0=0.050$ ,  $P_1=0.250$ ,  $\alpha=0.050$ ,  $\beta=0.200$

Stage 1 R1/N1	Stage 2 R2/N2	Stage 3 R3/N3	Ave N	Stage 1 Pet P0	Overall Pet P0	Alpha	Beta	Constraints Satisfied
2/16	2/16	2/16	16.00	0.000	0.000	0.043	0.197	Single Stage
0/12	1/15	2/16	13.55	0.833	0.869	0.043	0.199	Minimax
0/8	1/13	2/19	10.41	0.663	0.880	0.049	0.195	Optimum

#### Report Definitions

N1 is the sample size in the first stage.

R1 is the drug rejection number in the first stage.

N2 is the sample size in the first and second stages.

R2 is the drug rejection number in the second stage.

N3 is the combined sample size of all three stages.

R3 is the drug rejection number in the third stage.

Stage 1 PET P0 is the probability of early termination at the first stage.

Stage 2 PET P0 is the probability of early termination at the second stage.

Ave N is the average sample size if this design is repeated many times.

Alpha is the probability of rejecting that  $P \leq P_0$  when this is true.

Beta is the probability of rejecting that  $P \geq P_1$  when this is true.

P0 is the response proportion of a poor drug.

P1 is the response proportion of a good drug.

This report shows three designs. The first is the smallest single stage design. The second is the Minmax solution. This is the design with the smallest total sample size ( $N$ ). The third is the optimum design—the one that minimizes the average sample size.

Note that PASS matches the results of Chen.



## Chapter 135

# Post-Marketing Surveillance

## Introduction

Post-marketing surveillance refers to the search for adverse reactions to drugs that have been cleared for general use. Two types of study designs are often used: the cohort study and the case-control study. In a cohort design, a large group of treated patients are studied to determine the frequency of any adverse reactions. In a case-control study, patients who have experienced the adverse reaction are matched with other treated patients who have not.

## Technical Details

This section presents the formulas used to calculate sample size and power in four situations. The theory and formulas provided by Machin *et al.* (1997) are used.

### Design Type 1 - Cohort Study, No Background Incidence of Adverse Reactions

Let the anticipated incidence rate of adverse reactions be  $R0$ , the number of occurrences of a particular adverse reaction be  $A$ , the number of patients be  $N$ , and the probability that you will not find  $A$  reactions in the sample of  $N$  patients be  $\beta$ . If  $R0$  is small, the occurrence of an adverse reaction may be assumed to follow the Poisson distribution. If this is the case, the relationship among the above parameters is

$$\beta = \sum_{i=0}^{A-1} \frac{N^i (R0)^i e^{-N(R0)}}{i!}$$

Using numerical search techniques, *PASS* is able to solve any one of these parameters in terms of the others.

### Design Type 2 - Cohort Study, Known Background Incidence of Adverse Reactions

Let the anticipated incidence rate of adverse reactions be  $R0$ , let the additional incidence rate caused by the drug be  $D$ , and let the number of patients be  $N$ . For a given significance level  $\alpha$  and power  $1 - \beta$ , the relationship between these parameters is

$$z_{1-\beta} = \frac{D\sqrt{N} - z_{1-\alpha}\sqrt{R0}}{\sqrt{R0 + D}}$$

### Design Type 3 - Cohort Study, Unknown Background Incidence of Adverse Reactions

A control group is needed when the background incidence rate is not known. In post-marketing surveillance studies, the control group is usually made up of untreated individuals. Let the anticipated incidence rate of adverse reactions be  $R0$ , let the additional incidence rate caused by the drug be  $D$ , let the number of patients be  $N$ , and let the number of control patients for each treated patient be  $M$ . Thus the number of control patients is  $NM$ . For a given significance level  $\alpha$  and power  $1 - \beta$ , the relationship between these parameters is

$$z_{1-\beta} = \frac{D\sqrt{MN} - z_{1-\alpha}\sqrt{(M+1)R(1-R)}}{\sqrt{MR0(1-R0) + (R0+D)(1-R0-D)}}$$

where

$$R = \frac{R0 + M(R0 + D)}{1 + M}$$

### Design Type 4 - Matched Case-Control Study

A case-control design involves identifying a group of patients that have experienced the reaction of interest and then obtaining matched control patients that have not experienced the reaction.

Let the anticipated incidence rate of adverse reactions be  $R0$ , let the additional incidence rate caused by the drug be  $D$ , let the number of patients be  $N$ , and let the number of control patients matched with each treated patient be  $M$ . For a given significance level  $\alpha$  and power  $1 - \beta$ , the relationship between these parameters is

$$z_{1-\beta} = \frac{|R0 - \Omega|\sqrt{N} - z_{1-\alpha}\sqrt{\left(1 + \frac{1}{M}\right)\Pi(1-\Pi)}}{\sqrt{\frac{R0(1-R0)}{M} + (\Omega)(1-\Omega)}}$$

where

$$\Omega = \frac{R0 + D}{1 + D}$$

$$\Pi = \left(\frac{R0}{1 + D}\right)\left(M + \frac{\Omega}{R0}\right)$$

# 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.

## Design Type

This parameter specifies which of the four possible designs is to be analyzed.

Note that the other parameters have numbers in brackets, such as [234]. The numbers between the brackets indicate which of the four designs use that parameter.

The possible designs are:

### [1] Cohort - No Background Incidence

This is a cohort design in which the adverse effect does not occur except when caused by the drug.

### [2] Cohort - Known Background Incidence

This is a cohort design. The adverse effect can occur without being related to the drug. The incidence rate of the adverse effect is known.

### [3] Cohort - Unknown Background Incidence

This is a cohort design. Although the adverse effect can occur, its incidence rate is not known. A control group must be followed to determine the background incidence rate.

### [4] Matched Case-Control Study

One or more control patients is matched with each case patient. All of these patients are in the study. This is different from design type 3 in that the controls are matched with the cases. In design type 3, no matching is done.

## Find

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

## R0 (Incidence Rate)

This is the background incidence rate of the adverse reaction. This is the rate that occurs in the population without the drug being monitored. Since this is an incidence rate, and hence a proportion, it should be less than one. Also, this value plus  $D$  must be less than one.

The [234] reminds you that this parameter is used with design types 2, 3, and 4 (but not 1).

## D (Additional Incidence Rate)

This is the additional incidence rate of the adverse reaction that can be attributed to the drug being studied. Since this is a rate, it should not be greater than one. Also, this value plus  $RO$  should be less than one.

The [1234] reminds you that this parameter is used with all four design types.

## A (Number of Occurrences)

This is the number of occurrences of the adverse reaction of interest in the  $N$  patients being monitored. Sometimes, a single drug-related adverse reaction (such as death) might be enough to make the drug unacceptable. The acceptable range is one or greater.

The [1] reminds you that this parameter is only needed when the Design Type is 1.

## N (Patients)

This is the number of patients in the cohort being studied. In the case-control designs, this is the number of cases. For case-control studies, the total number of patients in the study is  $N(1+M)$  where  $M$  is the number of control patients per case patient.

The [1234] reminds you that this parameter is used with all four design types.

## M (Controls Per Case)

This is the number of control patients for each case patient. For case-control studies, the total number of patients in the study is  $N(1+M)$  where  $N$  is the number of case patients.

The [34] reminds you that this parameter is only used when the Design Type is 3 or 4.

## Alpha

This option specifies one or more values for the significance level--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 in your experimental situation.

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

The [234] reminds you that this parameter is not used when the Design Type is 1.

## One-Sided

This option lets you designate whether the test will be one-sided (checked) or two-sided (unchecked). When the two-sided box is indicated, the alpha value is simply divided by two.

## 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 was often used for beta. Recently, the value of 0.10 has become common. 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.

Note that the interpretation of Beta is a little different when the Design Type is 1.

## Options Tab

This tab sets a couple of options 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.

### Iterative Precision

When a search is made for the precision value, this is the cutoff value used to terminate the search. In most cases, a value of 0.0001 will be more than sufficient.

## Example1 - Calculating the Power

Suppose 1 in 10,000 people receiving a certain drug are expected to have an irregular heart beat. A researcher decides that if the irregular heart beat occurs in three patients, the drug will have to be withdrawn. Study the sample size necessary to achieve 99% probability of success.

In order to do this, sample sizes between 1000 and 21,000 will be considered.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Design Type.....	[1] Cohort - No Background Incidence
Find .....	Beta and Power
R0 .....	Ignored
D .....	0.0001
A.....	1 2 3
N .....	1000 to 21000 by 4000
M.....	Ignored
Alpha.....	Ignored
Beta.....	0.20
<b>Reports Tab</b>	
Incidence Rate Decimals .....	5

### Annotated Output

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

## Numeric Results

### Numeric Results (Cohort Study with No Background Incidence)

Power	Sample Size (N)	Additional Incidence Rate (D)	Number of Occurrences (A)	Beta
0.09516	1000	0.00010	1	0.90484
0.00468	1000	0.00010	2	0.99532
0.00015	1000	0.00010	3	0.99985
0.39347	5000	0.00010	1	0.60653
0.09020	5000	0.00010	2	0.90980
0.01439	5000	0.00010	3	0.98561
0.59343	9000	0.00010	1	0.40657
0.22752	9000	0.00010	2	0.77248
0.06286	9000	0.00010	3	0.93714
0.72747	13000	0.00010	1	0.27253
0.37318	13000	0.00010	2	0.62682
0.14289	13000	0.00010	3	0.85711
0.81732	17000	0.00010	1	0.18268
0.50675	17000	0.00010	2	0.49325
0.24278	17000	0.00010	3	0.75722
0.87754	21000	0.00010	1	0.12246
0.62039	21000	0.00010	2	0.37961
0.35037	21000	0.00010	3	0.64963

#### Report Definitions

Power is 1 - Beta.

N is the number of patients monitored.

D is the expected incidence rate of adverse reactions.

A is the number of adverse reactions.

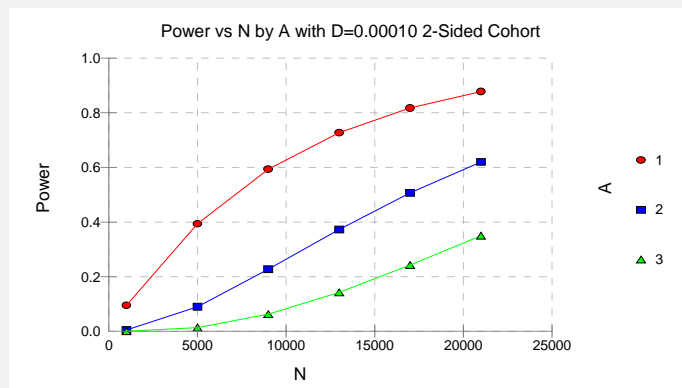
Beta is the probability that A reactions will not be found in the N patients.

#### Summary Statements

In a cohort study with no background incidence of a particular adverse reaction, the probability that 1 or more adverse reactions will not occur in a sample of 1000 patients with an anticipated incidence rate of 0.00 is 0.90484. The power of this study is 10%.

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

## Plot Section



This plot shows the power versus the sample size for three values of A.

## Example2 - Validation using Machin

Machin *et al.* (1997) page 147 give an example of a cohort design with no background incidence in which  $N$  is 30000, incidence is 0.0001, and  $A$  is 1 or 2. When  $A$  is 1, the power is 95%. When  $A$  is 2, the power is 80%.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Design Type .....	<b>[1] Cohort - No Background Incidence</b>
Find .....	<b>Beta and Power</b>
R0 .....	<i>Ignored</i>
D .....	<b>0.0001</b>
A .....	<b>1 2</b>
N .....	<b>30000</b>
M .....	<i>Ignored</i>
Alpha .....	<i>Ignored</i>
Beta .....	<i>Ignored</i>
<b>Reports Tab</b>	
Incidence Rate Decimals .....	<b>5</b>

### Annotated Output

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

Numeric Results (Cohort Study with No Background Incidence)				
Power	Sample Size (N)	Additional Incidence Rate (D)	Number of Occurrences (A)	Beta
0.95021	30000	0.00010	1	0.04979
0.80085	30000	0.00010	2	0.19915

*PASS* has calculated the same power values as did Machin *et al* (1997).



## Example3 - Validation using Machin

Machin *et al.* (1997) page 148 give an example of a cohort design with unknown background incidence in which  $N$  is 8500,  $R_0$  is 0.01,  $D$  is 0.005, and  $A$  is 1. The power is 90%.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Design Type .....	<b>[3] Cohort - Unknown Background Incidence</b>
Find .....	<b>Beta and Power</b>
$R_0$ .....	<b>0.01</b>
$D$ .....	<b>0.005</b>
$A$ .....	<b>1</b>
$N$ .....	<b>8500</b>
$M$ .....	<b>1</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored</i>
<b>Reports Tab</b>	
Incidence Rate Decimals .....	<b>5</b>

### Annotated Output

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

Numeric Results (Cohort Study with No Background Incidence)						
Power	Sample Size (N)	Controls Per Case (M)	Background Incidence Rate ( $R_0$ )	Additional Incidence Rate (D)	One-Sided Alpha	Beta
0.90136	8500	1	0.01000	0.00500	0.05000	0.09864

*PASS* has calculated the same power value as did Machin *et al* (1997).



## Chapter 150

# Inequality of Correlated Proportions

## Introduction

McNemar's test compares the proportions for two correlated dichotomous variables. These two variables may be two responses on a single individual or two responses from a matched pair (as in matched case-control studies).

This procedure is similar to the Matched Case-Control procedure also available in *PASS*. It differs from that procedure in three basic ways:

1. The results are calculated exactly using an unconditional formula rather than using conditional, normal approximations to the binomial.
2. It only deals with the case of a matched pair: one case and one control (the Matched Case-Control procedure lets you match several controls with each case).
3. It is based directly on the 2-by-2 contingency table.

To fix these ideas, consider the following fictitious data concerning the relationship between smoking and lung cancer. Suppose that a sample of  $N = 100$  cases of identical twins in which only one twin has lung cancer is selected for further study. The twin with lung cancer is the *case*. The other twin serves as the *control*. Each pair of twins is surveyed to determine if they smoke tobacco. The results are summarized in the following two-way table:

	<u>No Lung Cancer Twin (Control)</u>	
<u>Lung Cancer Twin (Case)</u>	<u>Smokes = Yes</u>	<u>Smokes = No</u>
<b>Smokes = Yes</b>	16	21
<b>Smokes = No</b>	4	59

There is a basic difference between this table and the more common two-way table. In the matched-paired case, the count represents the number of pairs, not the number of individuals.

The investigator wishes to compare the proportion of cases that smoke with the proportion of controls that smoke. The proportion of controls who smoke is  $(16+4)/100 = 0.20$ . The proportion of cases who smoke is  $(16+21)/100 = 0.37$ .

Dividing each of the entries in the table by  $N$  gives the proportions:

<u>No Lung Cancer Twin (Control)</u>			
<u>Lung Cancer Twin (Case)</u>	<u>Smokes = Yes</u>	<u>Smokes = No</u>	<u>Total</u>
Smokes = Yes	0.16	0.21	0.37
Smokes = No	0.04	0.59	0.63
Total	0.20	0.80	0.63

Symbolically, this table is represented as:

<u>No Lung Cancer Twin (Control)</u>			
<u>Lung Cancer Twin (Case)</u>	<u>Smokes = Yes</u>	<u>Smokes = No</u>	<u>Total</u>
Smokes = Yes	$P_{11}$	$P_{10}$	$P_t$
Smokes = No	$P_{01}$	$P_{00}$	$1 - P_t$
Total	$P_s$	$1 - P_s$	1

Formally, the hypothesis of interest is that  $P_t$  equals  $P_s$ . A little algebra shows that  $P_t = P_s$  is equivalent to  $P_{10} = P_{01}$ , since  $P_{11}$  is common to both. Thus, the null hypothesis of McNemar's test is  $P_{10} = P_{01}$  and the alternative is that they are unequal. The alternative hypothesis may be one-sided (such as  $P_{10} > P_{01}$ ) or two-sided ( $P_{10} \neq P_{01}$ ).

The null hypothesis may also be stated in terms of the odds ratio as  $OR = 1$ . The odds ratio is computed differently in the matched pairs case. The formula is:

$$OR = \frac{P_{12}}{P_{21}}$$

Notice that the values of  $P_{11}$  and  $P_{00}$  are not used in these hypotheses. It turns out that their individual values are not needed, but their sum is.

For this example, the odds ratio is computed as  $21/4 = 5.25$ .

## Technical Details

Consider the matched-pairs table again:

<u>Controls</u>			
<u>Cases</u>	<u>Yes</u>	<u>No</u>	<u>Total</u>
Yes	$P_{11}$	$P_{10}$	$P_t$
No	$P_{01}$	$P_{00}$	$1 - P_t$
Total	$P_s$	$1 - P_s$	1

Pairs with the same response from cases and controls (Yes-Yes and No-No) are called *concordant* pairs. Pairs with different responses (Yes-No and No-Yes) are called *discordant* pairs. McNemar's test statistic is the estimated odds ratio:

$$Mc = \frac{P_{10}}{P_{01}}$$

The sample size problem thus reduces to a study of how many Yes-No's and No-Yes's are needed. Once this has been determined, the overall sample size is found by estimating the proportion of discordant pairs and inflating the sample size appropriately.

Some power analysis programs follow an approximate procedure. Since the McNemar statistic follows the binomial probability distribution for a fixed number of discordant pairs, they use formulas that use the normal approximation to the binomial and then adjust the sample size depending on the proportion of discordant pairs,  $PD = P10 + P01$ . This is called the conditional procedure.

One such approximate formula is given by Machin, Campbell, Fayers, and Pinol (1997).

$$N_{pairs} = \frac{\left\{ z_{1-\alpha/s}(OR+1) + z_{1-\beta}\sqrt{(OR+1)^2 - (OR-1)^2 PD} \right\}^2}{(OR-1)^2 PD}$$

where  $s$  is the number of sides to the test (one or two),  $OR = \frac{P10}{P01}$ ,  $PD = P10 + P01$ , and alpha and beta are the usual error rate probabilities.

However, Schork and Williams (1980) published a formula which provides the exact results for the unconditional case. This is the formulation that is used in **PASS**.

$$Power = \sum_{R=r}^N \sum_{n_{12}=0}^{IR} \frac{N!}{(N-R)!n_{12}!(R-n_{12})!} (1-PD)^{N-R} \left( \frac{D+PD}{2} \right)^{n_{12}} \left( \frac{PD-D}{2} \right)^{R-n_{12}}$$

where

$$PD = P10 + P01$$

$$D = P10 - P01$$

$N$  is total of all four cells ( $N11 + N12 + N21 + N22$ )

$r$  is the smallest integer for which  $\left( \frac{1}{2} \right)^r \leq \alpha$

$IR$  is the largest integer such that  $\sum_{j=0}^{IR} \binom{R}{j} \left( \frac{1}{2} \right)^R \leq \alpha$

## Difference or Odds Ratio

The formula given above is parameterized in terms of the difference. This formula is also used when odds ratios are specified. The program simply converts the OR value into its corresponding D value.

## Discussion

The exact algorithm works for  $N < 2000$ . Above 2000, computing time goes up and the algorithm has numerical problems. **PASS** lets you select either the exact, or the approximate, solutions. We have found that the approximate solution tends to use a sample size that is about 10% less than the exact solution.

Because of the lengthy computer time necessary to compute the exact answer when  $N > 1500$ , we suggest that you use the approximate formula to begin with and then revert to the exact formula when you are ready for your final results. This is based on the value of  $N$  specified in the 'Use Approximations if  $N >$ ' box of the Options tab.

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers two procedures, which have different options. This section documents options that are common to both procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Odds Ratio*, *Difference*, *N*, *Alpha*, and *Beta* (or *Power*). Under most situations, you will select either *Beta* to calculate power or *N* to calculate sample size.

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. Two solutions can often be found when searching for the *Odds Ratio* or the *Difference*. You may specify the region in which you want the solution to be searched for. For example, you may search for an odds ratio either above or below one. Also note that the parameter selected must match the procedure you are using. For example, if you are searching for the odds ratio, you must be in the odds ratio window.

Note that the value selected here always appears as the vertical axis on the charts.

## Proportion Discordant (P10+01)

This is the proportion of discordant pairs ( $P10 + P01$ ). This value will be difficult to specify unless you have previous studies that give you some idea of what to expect. When you have no idea, Machin, Campbell, Fayers, and Pinol (1997) suggest the following strategy. Estimate values of  $P_t$  and  $P_s$ . Calculate the proportion of discordant pairs using the approximation

$$PD = P_t(1 - P_s) + P_s(1 - P_t)$$

This approximation assumes that the two responses are independent in each subject, which will usually not be true. However, it may be the only way of determining a ball park value for this parameter.

## Use Approximations if N >

Below this value of  $N$ , the exact power calculation formula based on the binomial is used. Above this value of  $N$ , the approximate formula based on the normal approximation to the binomial is used. The exact formula suffers from numerical problems when  $N$  is greater than 2000. On the other hand, the approximate formula tends to underestimate the  $N$  necessary to achieve a certain beta value by about 5%.

You control which formula is used by setting this value.

## Alternative Hypothesis

Specify whether the test is one-sided or two-sided. A one-sided hypothesis uses an inequality as in  $P_{10} > P_{01}$  or Odds Ratio  $> 1$ . A two-sided hypothesis states that the proportions 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.

## N (Number of Pairs)

Enter a value (or range of values) for the sample size,  $N$ , the number of pairs in the study.

You may enter a range of values such as *100,200,300* or *200 to 400 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 conclude equivalence or inferiority.

## 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  $H_0$ . Power is defined as 1-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.

Values must be between 0 and 1. Values of 0.10 or 0.20 are often used for beta. However, you should pick a value for beta that represents the risk of a type-II error that you are willing to take.

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

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P_{10} - P_{01}$ .

## Difference (P10-P01)

The difference,  $P_{10} - P_{01}$ , is a popular parameter to specify because it comes directly from the alternative hypothesis that  $P_t$  not equal  $P_s$ . If  $P_t$  and  $P_s$  are not equal, the obvious question is, by how much? The answer is, by  $P_{10} - P_{01}$  since  $P_t - P_s = (P_{11} + P_{10}) - (P_{11} + P_{01}) = P_{10} - P_{01}$ . Hence, this is a value that may easily be set.

The range of values is between -1 and 1. You may enter a list of value list *0.1, 0.15, 0.2* or *0.05 to 0.20 by 0.05*.

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratio,  $P10 / P01$ .

### Odds Ratio ( $P10/P01$ )

The odds ratio is a popular parameter to specify because of its simple interpretability and close relationship to the relative risk. An odds ratio of 2.0 means that the odds of the numerator variable is twice the odds of the denominator variable. Note that several values of  $P10$  and  $P01$  can yield the same odds ratio. For example,  $0.2/0.1$  and  $0.4/0.2$  both have an odds ratio of 2.0, but are based on very different values of  $P10$  and  $P01$ . Under the null hypothesis, the odds ratio is one. Only positive values are allowed.

You may enter a list of values like *1.5,2.0,3.0* or *1.2 to 2.4 by 0.2*.

## Options Tab

This tab sets a couple of options 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

This example will show how to calculate the power of a retrospective study for several sample sizes and odds ratio values. Suppose that a matched case-control study is to be run in which the odds ratios might be 1.5, 2.5, or 3.5;  $PD$  is 0.3,  $N = 50$  to 300 by 50; alpha is 0.05; and beta is to be calculated.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the odds ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Odds Ratio .....	<b>1.5 2.0 2.5</b>
Proportion Discordant .....	<b>0.3</b>
Use Approximations if N>.....	<b>1500</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N (Number of Pairs) .....	<b>50 to 200 by 50</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>Calculated</b>

# Annotated Output

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

## Numeric Results

### Numeric Results for Two-Sided Test

Power	N	P10	P01	Difference (P10-P01)	Proportion Discordant	Odds Ratio	Alpha	Beta
0.07467	50	0.180	0.120	0.060	0.300	1.500	0.05000	0.92533
0.14214	100	0.180	0.120	0.060	0.300	1.500	0.05000	0.85786
0.22451	150	0.180	0.120	0.060	0.300	1.500	0.05000	0.77549
0.29790	200	0.180	0.120	0.060	0.300	1.500	0.05000	0.70210
0.17849	50	0.200	0.100	0.100	0.300	2.000	0.05000	0.82151
0.37305	100	0.200	0.100	0.100	0.300	2.000	0.05000	0.62695
0.56457	150	0.200	0.100	0.100	0.300	2.000	0.05000	0.43543
0.70340	200	0.200	0.100	0.100	0.300	2.000	0.05000	0.29660
0.29601	50	0.214	0.086	0.129	0.300	2.500	0.05000	0.70399
0.59122	100	0.214	0.086	0.129	0.300	2.500	0.05000	0.40878
0.80105	150	0.214	0.086	0.129	0.300	2.500	0.05000	0.19895
0.90751	200	0.214	0.086	0.129	0.300	2.500	0.05000	0.09249

### Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.

N is the number of pairs in the sample.

P10 is the proportion of pairs in cell 1,2 of the 2x2 table.

P01 is the proportion of pairs in cell 2,1 of the 2x2 table.

Difference is the difference between proportions parameter under the alternative hypothesis.

Proportion Discordant' is the total of P10 and P01.

Odds Ratio is the value of this parameter under the alternative hypothesis.

Alpha is the probability of rejecting a true null hypothesis. It should be small.

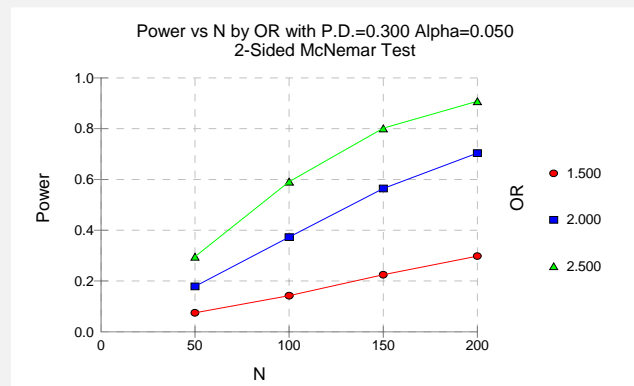
Beta is the probability of accepting a false null hypothesis. It should be small.

### Summary Statements

A sample size of 50 pairs achieves 7% power to detect an odds ratio of 1.500 using a two-sided McNemar test with a significance level of 0.05000. The odds ratio is equivalent to a difference between two paired proportions of 0.060 which occurs when the proportion in cell 1,2 is 0.180 and the proportion in cell 2,1 is 0.120. The proportion of discordant pairs is 0.300.

This report shows the power for each of the scenarios.

## Plot Section



This plot shows the power versus the sample size for the three odds ratios.

## Example2 - Validation using Schork and Williams

Schork and Williams (1980) page 354 present a table of sample sizes for various combinations of the other parameters. When the difference is 0.2, the proportion discordant is 0.7, the power is 80%, and the one-sided significance level is 0.025, the sample size is 144.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Difference .....	<b>0.2</b>
Proportion Discordant .....	<b>0.7</b>
Alternative Hypothesis .....	<b>One-Sided</b>
N (Number of Pairs) .....	<b>144</b>
Alpha .....	<b>0.025</b>
Beta .....	<i>Ignored</i>

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

Numeric Results for Two-Sided Test								
Power	N	P10	P01	Difference (P10-P01)	Proportion Discordant	Odds Ratio	Alpha	Beta
0.80092	144	0.450	0.250	0.200	0.700	1.800	0.02500	0.19908

*PASS* has also found the power to be about 80%.



## Chapter 155

# Matched Case-Control

## Introduction

A 2-by- $M$  case-control study investigates a risk factor relevant to the development of a disease. A population of *case* patients with a disease and *control* patients without the disease is considered. Some of these patients have had exposure to a risk factor of interest. A random sample of  $N$  case patients is selected. Patients are stratified by the levels of a confounding variable (such as age, gender, etc.). For each selected case patient, a random sample of  $M$  matched control patients is drawn from the same strata (group). An estimate of the odds ratio,  $OR$ , of developing the disease in exposed and unexposed patients who have equal values of the confounding variable is desired. This odds ratio is assumed to be constant across all levels of the confounding variables.

To fix these ideas, consider the following fictitious data concerning the relationship between smoking and lung cancer. Suppose that a sample of  $N = 100$  cases of identical twins in which only one twin has lung cancer is selected. The second twin serves as the control. Each pair of twins is surveyed to determine if either, both, or none smoke tobacco. The results are summarized in the following two-way table:

	<u>No Lung Cancer Twin (Control)</u>	
<u>Lung Cancer Twin (Case)</u>	<u>Smokes = Yes</u>	<u>Smokes = No</u>
<b>Smokes = Yes</b>	16	21
<b>Smokes = No</b>	4	59

Note that the values in this table are counts of pairs, not individuals. The proportion of controls who smoke is  $(16+4)/100 = 0.20$ . The proportion of cases who smoke is  $(16+21)/100 = 0.37$ . The odds ratio is  $21/4 = 5.25$ . That is, the twin who smoked is 5.25 times more likely to have lung cancer than the twin who did not.

This procedure is similar to the McNemar procedure also available in *PASS*. It differs from that procedure in three basic ways:

1. The results are based on the normal approximation to the binomial.
2. This procedure lets you have multiple controls for each case. The McNemar procedure only allows one control per case.
3. The input parameters are different.

## Technical Details

The following results summarize the article by Dupont (1988) upon which this module is based. The probabilities that the data fall into various categories are:

1. The probability that a case patient was exposed to the risk factor is  $p_1$ .
2. The probability that a control patient was exposed to the risk factor is  $p_0$ .
3. The probability that a case patient was not exposed to the risk factor is  $q_1 = 1 - p_1$ .
4. The probability that a control patient was not exposed to the risk factor is  $q_0 = 1 - p_0$ .

The odds ratio,  $OR$ , is defined as

$$OR = \frac{p_1 / q_1}{p_0 / q_0}$$

Assume that you use a  $\chi^2$  test for the null hypothesis that  $OR = 1$ , that is, that  $p_0 = p_1$ . Such a test is given by Breslow and Day (1980).

Let  $x_k = 1$  or  $0$  if the  $k^{th}$  sampled case patient was or was not exposed, respectively. Let  $y_k = 1$  or  $0$  if the corresponding first matched control patient was or was not exposed. Let  $p_{ij} = \Pr(x_k = i \text{ and } y_k = j)$ . For example,  $p_{10}$  is the probability that the case patient was exposed to the risk factor while the corresponding first control patient was not. The relationships between these probabilities are

$$p_1 = p_{11} + p_{10}$$

and

$$p_0 = p_{11} + p_{01}$$

Define  $\phi$  to be the correlation between  $x_k$  and  $y_k$ . It can be shown that

$$\phi = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{p_1q_1p_0q_0}}$$

A little algebra will show that

$$p_{11} = p_1p_0 + \phi\sqrt{p_1q_1p_0q_0}$$

$$p_{10} = p_1q_0 - \phi\sqrt{p_1q_1p_0q_0}$$

$$p_{01} = q_1p_0 - \phi\sqrt{p_1q_1p_0q_0}$$

$$p_{00} = q_1q_0 + \phi\sqrt{p_1q_1p_0q_0}$$

$$p_{0+} = \frac{p_{11}}{p_1}$$

$$p_{0-} = \frac{p_{01}}{q_1}$$

$$q_{0+} = 1 - p_{0+}$$

$$q_{0\cdot} = 1 - p_{0\cdot}$$

and

$$t_k = p_1 \binom{M}{k-1} p_{0+}^{k-1} q_{0+}^{M-k+1} + q_1 \binom{M}{k} p_{0\cdot}^k q_{0\cdot}^{M-k}; \quad k = 1, \dots, M$$

Let  $n_{ij}$  represent the number of matched sets of subjects in which the case patient was ( $i = 1$ ) or was not ( $i = 0$ ) exposed and  $j$  of the  $M$  control subjects were exposed. Let

$$y = \sum_{m=1}^M n_{1,m-1}$$

be the number of discordant sets in which the case patient was exposed and let

$$T_m = n_{1,m-1} + n_{0,m}$$

be the number of sets in which  $m$  subjects were exposed. The expected value of  $T_m$  is  $Nt_m$ .

Let

$$e_{OR} = \sum_{m=1}^M \frac{mt_m(OR)}{m(OR) + M - m + 1}$$

$$v_{OR} = \sum_{m=1}^M \frac{mt_m(OR)(M - m + 1)}{(m(OR) + M - m + 1)^2}$$

Dupont (1988) provides the following formula relating  $\alpha$ ,  $\beta$ ,  $p_0$ ,  $\phi$ ,  $OR$ ,  $N$ , and  $M$ .

$$1 - \beta = \Phi\left(\frac{\sqrt{N}(e_1 - e_{OR}) - z_{\alpha/2}\sqrt{v_1}}{\sqrt{v_{OR}}}\right) + 1 - \Phi\left(\frac{\sqrt{N}(e_1 - e_{OR}) + z_{\alpha/2}\sqrt{v_1}}{\sqrt{v_{OR}}}\right)$$

This equation may be used to make power and sample size calculations.

## Estimating the Sample Control Exposure Probability

To calculate power and sample size, a value for the probability that a sample control patient was exposed to the risk factor ( $p_0$ ) must be estimated. Remember that the control sample is not a random sample of the population. Instead, it is matched to a random sample of case patients. Hence, the sample does not necessarily provide an unbiased estimate of  $p_0$ . Care should be taken to provide an accurate estimate of the probability that a matched control patient was exposed, not the probability that someone was exposed in the general population. However, when there is little association between the confounding (matching) variable and the exposure variable in the control population, the baseline probability of the exposure variable may be used.

## Estimating the Correlation, $\phi$

Previous matched 2-by-2 contingency tables can be used to estimate  $\phi$  using the relationship

$$\phi = \sqrt{\frac{\chi_u^2}{N}}$$

where

$$\chi_u^2 = \frac{N(n_{00}n_{11} - n_{01}n_{10})^2}{n_{0\bullet}n_{1\bullet}n_{\bullet 0}n_{\bullet 1}}$$

When no previous data are available about  $\phi$ , Dupont (1988) suggests using a value of 0.2 rather than 0.

## 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 *Odds Ratio*, *Probability Exposed*, *Correlation*, *N*, *M*, *Alpha*, and *Beta*. 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.

### OR (Odds Ratio)

This option sets the value of the odds ratio, *OR*, which is the ratio of the disease odds of individuals exposed to the risk factor to the disease odds of individuals not exposed to the risk factor.

For example, an odds ratio of 2.0 means that subjects exposed to the risk factor are twice as likely of developing the disease as are unexposed subjects.

A value greater than one is usually used. The value must be greater than zero. The null hypothesis is that the odds ratio is one. You should use a value that will be of interest to others, such as 1.5 or 2.0.

You can enter a list of values.



## P0 (Prob Exposed)

Specify the value of  $p_0$ , the probability that a sample control patient is exposed to the risk factor. If the matching variables are independent of the risk factor, you can use the baseline rate of exposure to the risk factor in the general population.

Since this is a probability, it must be between zero and one.

You can enter a list of values separated by commas or blanks.

## Phi (Correlation)

This is the correlation for exposure between a case subject and the first of the corresponding control subjects. A value of zero here indicates independence between exposure rates for case and controls. Often, assuming complete independence is unrealistic, so when no other information is available, Dupont (1988) suggests using a value of 0.2.

Correlations can range between -1 and 1. However, only positive correlations should be used.

## N (Case Patients)

The number of case patients to be sampled. This is the number of pairs. Since there are  $M$  controls for each case patient, the total number of patients is  $N(M + 1)$ . Values of one or greater are allowed here.

## M (Controls Per Case)

The number of control patients matched with each case patient. When  $M$  is one, McNemar's test is used to analyze the data.

The total number of patients in the study is  $N(M + 1)$ . Values of one or greater are allowed here.

## 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. In this case, the test is two-sided.

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. 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 when in fact it is false.

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

# Example1 - Calculating Power

## Problem Statement

This example will show how to calculate the power of a retrospective study for several sample sizes and odds ratios.

Suppose that a matched case-control study is to be run in which the odds ratios of interest are 1.5, 2.5, or 3.5,  $P0 = 0.6$ , correlation = 0.2,  $M = 1$ ,  $N = 25\ 50\ 100\ 150\ 200$ , alpha = 0.05, and power is to be found.

## Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
OR.....	<b>1.5 2.5 3.5</b>
P0.....	<b>0.6</b>
Correlation .....	<b>0.2</b>
N .....	<b>25 50 100 150 200</b>
M .....	<b>1</b>
Alpha.....	<b>0.05</b>
Beta.....	<b><i>Ignored since this is the Find Setting</i></b>

## Annotated Output

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

## Numeric Results

### Numeric Results

Power	Cases (N)	Controls Per Case (M)	Odds Ratio (OR)	Probability Exposed (P0)	Correlation (Phi)	Alpha	Beta
0.08863	25	1	1.50	0.60000	0.20000	0.05000	0.91137
0.13364	50	1	1.50	0.60000	0.20000	0.05000	0.86636
0.22622	100	1	1.50	0.60000	0.20000	0.05000	0.77378
0.31832	150	1	1.50	0.60000	0.20000	0.05000	0.68168
0.40652	200	1	1.50	0.60000	0.20000	0.05000	0.59348
0.23067	25	1	2.50	0.60000	0.20000	0.05000	0.76933
0.44278	50	1	2.50	0.60000	0.20000	0.05000	0.55722
0.75646	100	1	2.50	0.60000	0.20000	0.05000	0.24354
0.90966	150	1	2.50	0.60000	0.20000	0.05000	0.09034
0.97004	200	1	2.50	0.60000	0.20000	0.05000	0.02996
0.36379	25	1	3.50	0.60000	0.20000	0.05000	0.63621
0.68570	50	1	3.50	0.60000	0.20000	0.05000	0.31430
0.95159	100	1	3.50	0.60000	0.20000	0.05000	0.04841
0.99482	150	1	3.50	0.60000	0.20000	0.05000	0.00518
0.99956	200	1	3.50	0.60000	0.20000	0.05000	0.00044

### Report Definitions

Power is the probability of rejecting a false null hypothesis.

N is the size of the sample drawn from the treatment (case) group.

M is the number of matching control patients drawn for each case patient.

OR is the odds ratio of for subjects exposed to the risk factor.

P0 is the probability of exposure among sampled control patients.

Phi is the correlation of exposure between matched individuals.

Alpha is the probability of rejecting a true null hypothesis.

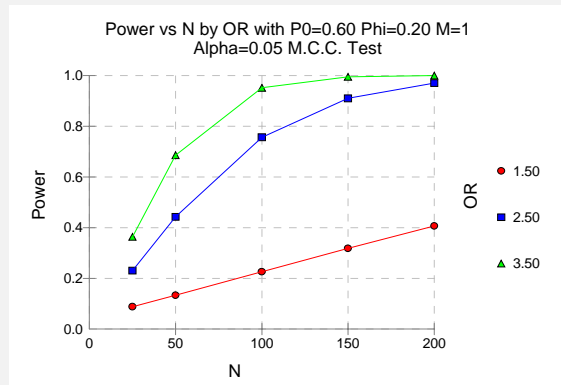
Beta is the probability of accepting a false null hypothesis.

### Summary Statements

In a matched case-control study, the probability of exposure among sampled control patients is 0.60000 and the correlation coefficient for exposure between matched case and control patients is 0.20000. A sample of 25 case patients is obtained. For each case patient, a matching sample of 1 control patient(s) is also obtained. This sample of 50 patients achieves 9% power to detect an odds ratio of 1.50 versus the alternative of equal odds using a Chi-Square test with a 0.05000 significance level.

This report shows the power for each of the scenarios.

## Plot Section



This plot shows the power versus the sample size for the three odds ratios.

## Example2 - Calculating Sample Size

Suppose that a matched case-control study is planned to study the relationship between smoking and a certain kind of cancer. Researchers want to have a sample large enough to detect an odds ratio of 2.0. During the power analysis, the researchers also want to calculate the required sample size for odds ratios of 1.5 and 2.5.

The probability that a sampled control (non-cancer) patient smokes is estimated at 0.3. The correlation of smoking between cases and controls is 0.2. The researchers want samples sizes large enough to achieve 80% power at the 0.05 significance levels. In an effort to reduce the number of cancer patients that must be enrolled, the researchers want to try several values for the number of controls per case between 1 and 20.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N (Cases)</b>
OR.....	<b>1.5 2.0 2.5</b>
P0.....	<b>0.3</b>
Correlation .....	<b>0.2</b>
N .....	<b><i>Ignored since this is the Find Setting</i></b>
M.....	<b>1 2 3 4 5 10 20</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>

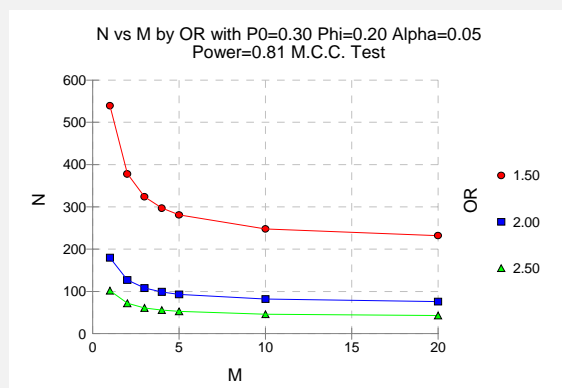
### Annotated Output

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

## Numeric Results

### Numeric Results

Power	Cases (N)	Controls Per Case (M)	Odds Ratio (OR)	Probability Exposed (P0)	Correlation (Phi)	Alpha	Beta
0.80064	539	1	1.50	0.30000	0.20000	0.05000	0.19936
0.80021	378	2	1.50	0.30000	0.20000	0.05000	0.19979
0.80025	324	3	1.50	0.30000	0.20000	0.05000	0.19975
0.80038	297	4	1.50	0.30000	0.20000	0.05000	0.19962
0.80076	281	5	1.50	0.30000	0.20000	0.05000	0.19924
0.80009	248	10	1.50	0.30000	0.20000	0.05000	0.19991
0.80044	232	20	1.50	0.30000	0.20000	0.05000	0.19956
0.80100	180	1	2.00	0.30000	0.20000	0.05000	0.19900
0.80249	127	2	2.00	0.30000	0.20000	0.05000	0.19751
0.80077	108	3	2.00	0.30000	0.20000	0.05000	0.19923
0.80220	99	4	2.00	0.30000	0.20000	0.05000	0.19780
0.80087	93	5	2.00	0.30000	0.20000	0.05000	0.19913
0.80261	82	10	2.00	0.30000	0.20000	0.05000	0.19739
0.80125	76	20	2.00	0.30000	0.20000	0.05000	0.19875
0.80156	102	1	2.50	0.30000	0.20000	0.05000	0.19844
0.80262	72	2	2.50	0.30000	0.20000	0.05000	0.19738
0.80028	61	3	2.50	0.30000	0.20000	0.05000	0.19972
0.80327	56	4	2.50	0.30000	0.20000	0.05000	0.19673
0.80559	53	5	2.50	0.30000	0.20000	0.05000	0.19441
0.80351	46	10	2.50	0.30000	0.20000	0.05000	0.19649
0.80672	43	20	2.50	0.30000	0.20000	0.05000	0.19328



This report shows the sample size for each of the scenarios. Notice that the required number of cancer patients ( $N$ ) drops off drastically as more controls are added. However, using more than five controls seems to only moderately reduce the sample size necessary sample size.

Also notice that the difference in sample size is much larger when moving from an odds ratio of 2.0 to 1.5 than from 2.5 to 2.0.

## Example3 - Validation using Dupont

The formulas used in this module were given in Dupont (1988). He gives an example on page 1164 of the article in which  $P0$  is 0.6,  $\Phi$  is 0.2,  $OR$  is 3.0,  $\alpha$  is 0.05, and  $\beta$  is 0.2. Dupont finds the sample size for  $M = 1$  to be 80 and for  $M = 3$  to be 50.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N (Cases)</b>
OR.....	<b>3.0</b>
P0.....	<b>0.6</b>
Correlation .....	<b>0.2</b>
N .....	<b><i>Ignored since this is the Find Setting</i></b>
M .....	<b>1 3</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

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

#### Numeric Results

Numeric Results								
Power	Cases Controls Per		Odds Ratio (OR)	Probability Exposed (P0)	Correlation (Phi)	Alpha	Beta	
	(N)	Case (M)						
0.80149	80	1	3.00	0.60000	0.20000	0.05000	0.19851	
0.80052	50	3	3.00	0.60000	0.20000	0.05000	0.19948	

Note that values of 80 and 50 for  $N$  agree exactly with Dupont.

## Chapter 160

# Non-Inferiority of Correlated Proportions

These modules compute power and sample size for non-inferiority tests in which two dichotomous responses are measured on each subject. When one is interested in showing that the true proportions are different, the data are often analyzed with McNemar's test. However, we are interested in showing non-inferiority rather than difference. For example, suppose a diagnostic procedure is accurate, but is expensive to apply or has serious side effects. A replacement procedure is sought which is no less accurate, but is less expensive or has fewer side effects. In this case, we are not interested in showing that the two diagnostic procedures are different, but rather that the second is no worse than the first. *Non-inferiority tests* were designed for this situation.

These tests are often divided into two categories: *equivalence* (two-sided) tests and *non-inferiority* (one-sided) tests. Here, the term *equivalence tests* means that we want to show that two diagnostic procedures are equivalent—that is, their accuracy is about the same. This requires a two-sided hypothesis test. On the other hand, *non-inferiority tests* are used when we want to show that a new (experimental) procedure is no worse than the existing (reference or gold-standard) one. This requires a one-sided hypothesis test. The procedures discussed in this chapter deal with the non-inferiority (one-sided) case.

## Technical Details

The results of a study in which two dichotomous responses are measured on each subject can be displayed in a 2-by-2 table in which one response is shown across the columns and the other is shown down the rows. In the discussion to follow, the columns of the table represent the standard (reference or control) response and the rows represent the treatment (experimental) response. The outcome probabilities can be classified into the following table.

Experimental Diagnosis	Standard Yes	Diagnosis No	Total
Yes	$p_{11}$	$p_{10}$	$P_T$
No	$p_{01}$	$p_{00}$	$1 - P_T$
Total	$P_S$	$1 - P_S$	1

In this table,  $p_{ij} = p_{Treatment, Standard}$ . That is, the first subscript represents the response of the new, experimental procedure while the second subscript represents the response of the standard procedure. Thus,  $p_{01}$  represents the proportion having a negative treatment response and a positive standard response.

To aid in interpretation, analysts have developed a few proportions that summarize the table. Three of the most popular ratios are *sensitivity*, *specificity*, and *prevalence*.

### Sensitivity

Sensitivity is the proportion of subjects with a positive standard response who also have a positive experimental response. In terms of proportions from a 2-by-2 table,

$$\text{Sensitivity} = p_{11} / (p_{01} + p_{11}) = p_{11} / P_S$$

### Specificity

Specificity is the proportion of subjects with a negative standard response who also have a negative experimental response. In terms of proportions from a 2-by-2 table,

$$\text{Specificity} = p_{00} / (p_{10} + p_{00})$$

### Prevalence

Prevalence is the overall proportion of individuals with the disease (or feature of interest). In terms of proportions from a 2-by-2 table,

$$\text{Prevalence} = P_S.$$



The outcome counts from a sample of  $n$  subjects can be classified into the following table.

Experimental Diagnosis	Standard Yes	Diagnosis No	Total
Yes	$n_{11}$	$n_{10}$	$n_T$
No	$n_{01}$	$n_{00}$	$n - n_T$
Total	$n_S$	$n - n_S$	$n$

Note that  $n_{11} + n_{00}$  is the number of matches (*concordant pairs*) and  $n_{10} + n_{01}$  is the number of *discordant pairs*.

The hypothesis of interest concerns the two marginal probabilities  $P_T$  and  $P_S$ .  $P_S$  represents the accuracy or success of the standard test and  $P_T$  represents the accuracy or success of the new, experimental test. Non-inferiority is defined in terms of either the difference,  $D = P_T - P_S$ , or the relative risk ratio,  $R = P_T / P_S$ , of these two proportions. The choice between  $D$  and  $R$  will usually lead to different sample sizes to achieve the same power.

## Non-Inferiority Hypotheses using Difference

This section is based on Liu, Hsueh, Hsieh and Chen (2002). Refer to that paper for complete details.

The null and alternative hypotheses of non-inferiority in terms of the difference are

$$H_0: P_T - P_S \leq -D_E \text{ versus } H_1: P_T - P_S > -D_E$$

To demonstrate non-inferiority, one desires to reject the null hypothesis and thus conclude that the experimental treatment is not worse than the standard by as much or more than  $D_E$ . In the context of the preceding statement,  $D_E$  is defined to be positive. The choice of an appropriate  $D_E$  may be difficult. It should be clinically meaningful so that clinicians are willing to conclude that the experimental treatment is acceptable if the difference is less than  $D_E$ . From a statistical perspective,  $D_E$  should be less than the effect, if known, of the standard treatment compared to placebo.

Liu et al. (2002) discuss the RMLE-based (score) method for constructing these confidence intervals. This method is based on (developed by, described by) Nam (1997).

## Asymptotic Tests

An asymptotic test is given by

$$Z_L = \frac{\hat{D} + D_E}{\hat{\sigma}} = \frac{c + nD_E}{\sqrt{d - n\hat{D}^2}} \geq z_\alpha$$

where

$$\hat{D} = \frac{n_T}{n} - \frac{n_S}{n} = \frac{n_{10}}{n} - \frac{n_{01}}{n}$$

$$d = n_{10} + n_{01}$$

$$c = n_{10} - n_{01}$$

and  $z_\alpha$  is the standard normal deviate having  $\alpha$  in the right tail.

An estimate of  $\hat{\sigma}$  based on the RMLE-based (score) procedure of Nam (1997) uses the estimates

$$\tilde{\sigma}_L = \sqrt{\frac{\tilde{p}_{L,10} + \tilde{p}_{L,01} - D_E^2}{n}}$$

where

$$\tilde{p}_{L,10} = \frac{-\tilde{a}_L + \sqrt{\tilde{a}_L^2 - 8\tilde{b}_L}}{4}$$

$$\tilde{p}_{L,01} = \tilde{p}_{L,10} - D_E$$

$$\tilde{a}_{L,01} = -\hat{D}(1 - D_E) - 2(\hat{p}_{01} + D_E)$$

$$\tilde{b}_{L,01} = D_E(1 + D_E)\hat{p}_{01}$$

## Power Formula

The power when the actual difference is  $D_A$  can be evaluated exactly using the multinomial distribution. However, when the sample size is above a user-set level, we use a normal approximation to this distribution which leads to

$$1 - \beta(D_A) = \begin{cases} 1 - \Phi(c_L) & \text{if } D_A > -D_E \\ 0 & \text{otherwise} \end{cases}$$

where

$$c_L = \frac{-D_A}{\sigma} - \frac{D_E}{\sigma} + \frac{z_\alpha}{w_L}$$

$$\sigma = \sqrt{\frac{p_{01} + p_{10} - D_A}{n}}$$

$$w_L = \sqrt{\frac{2p_{01} + D_A - D_A^2}{2\bar{p}_{L,01} - D_E - D_E^2}}$$

$$\bar{p}_{L,01} = \frac{-a_L + \sqrt{a_L^2 - 8b_L}}{4}$$

$$a_L = -D_A(1 - D_E) - 2(p_{01} + D_E)$$

$$b_L = D_E(1 + D_E)p_{01}$$

## Non-Inferiority Hypotheses using Ratios

The following is based on Nam and Blackwelder (2002). We refer you to this paper for the complete details of which we will only provide a brief summary here.

When  $R_E < 1$ , the statistical hypotheses of non-equivalence are

$$H_0: P_T / P_S \leq R_E \text{ versus } H_1: P_T / P_S > R_E$$

### Test Statistics

The test statistic for an asymptotic test based on constrained maximum likelihood for large  $n$  is given by

$$Z(R_E) = \sqrt{\frac{n(\hat{P}_T - R_E \hat{P}_S)}{R_E(\tilde{p}_{10} + \tilde{p}_{01})}}$$

where

$$\tilde{p}_{10} = \frac{-\hat{P}_T + R_E^2(\hat{P}_S + 2\hat{p}_{10}) + \sqrt{(\hat{P}_T - R_E^2 \hat{P}_S)^2 + 4R_E^2 \hat{p}_{10} \hat{p}_{01}}}{2R_E(R_E + 1)}$$

$$\tilde{p}_{01} = R_E \tilde{p}_{10} - (R_E - 1)(1 - \hat{p}_{00})$$

$$\hat{p}_{01} = \frac{n_{01}}{n}, \hat{p}_{10} = \frac{n_{10}}{n}, \hat{P}_T = \frac{n_{10} + n_{11}}{n}, \hat{P}_S = \frac{n_{01} + n_{11}}{n}$$

## Power Formula

The power when the true value of the relative risk ratio is  $R_E$  can be evaluated exactly using the multinomial distribution. When  $n$  is large, we use a normal approximation to the multinomial distribution which leads to

$$\beta(R_A) = \Phi(c_U)$$

where

$$c_U = \frac{z_{1-\alpha} \sqrt{\bar{V}_0(T_0)} - E_1(T_0)}{\sqrt{V_1(T_0)}}$$

$$\bar{V}_0(T_0) = \frac{R_E(\bar{p}_{10} + \bar{p}_{01})}{n}$$

$$E_1(T_0) = (R_A - R_E)P_S$$

$$V_1(T_0) = \frac{(R_A + R_E^2)P_S - 2R_E p_{11} - (R_A - R_E)^2 P_S^2}{n}$$

$$\bar{p}_{10} = \frac{-P_T + R_E^2(P_S + 2p_{10}) + \sqrt{(P_T - R_E^2 P_S)^2 + 4R_E^2 p_{10} p_{01}}}{2R_E(R_E + 1)}$$

$$\bar{p}_{01} = R_E \bar{p}_{10} - (R_E - 1)(1 - p_{00})$$

## Nuisance Parameter

Unfortunately, the 2-by-2 table includes four parameters  $p_{11}$ ,  $p_{10}$ ,  $p_{01}$ , and  $p_{00}$ , but the power specifications above only specify two parameters:  $P_S$  and  $D_A$  or  $R_A$ . A third parameter is defined implicitly since the sum of the four parameters is one. One parameter, known as a nuisance parameter, remains unaccounted for. This parameter must be addressed to fully specify the problem. This fourth parameter can be specified by specifying any one of the following:

$p_{11}$ ,  $p_{10}$ ,  $p_{01}$ ,  $p_{00}$ ,  $p_{10} + p_{01}$ ,  $p_{11} + p_{00}$ , or the sensitivity of the experimental response,  $p_{11} / P_S$ .

It may be difficult to specify a reasonable value for the nuisance parameter since its value may not be even approximately known until after the study is conducted. Because of this, we suggest that you calculate power or sample size for a range of values of the nuisance parameter. This will allow you to determine how sensitive the results are to its value.

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers two procedures which have different options. This section documents options that are common to both procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Beta and Power* or *N*.

### Ps (Standard Proportion)

This is the proportion of yes's (or successes),  $P_s$ , when subjects received the standard treatment. This value or a good estimate is often available from previous studies.

You may enter a set of values separated by blank spaces. For example, you could enter '0.50 0.60 0.70'. Values between, but not including, 0 and 1 are permitted.

### Max N Using Exact Power

Specify the maximum value of N (sample size) for which you would like an exact power calculation based on the multinomial distribution. Sample Sizes greater than this value will use the asymptotic approximation given in the documentation. The exact calculation of the multinomial distribution becomes very time consuming for  $N > 200$ . For most cases, when  $N > 200$ , the difference between the exact and approximate calculations is small. For  $N > 200$ , the length of time needed to calculate the exact answer may become prohibitive. However, as the speed of computers increases, it will become faster and easier to calculate the exact power for larger values of N.

If you want all calculations to use exact results, enter '1000' here.

If you want all calculations to use the quick approximations, enter '1' here.

### N (Sample Size)

Enter a value for the sample size. This value must be greater than two. You may enter a range of values such as *10 to 100 by 10*.

### 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 conclude equivalence.

## 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 conclude equivalence when in fact it is true.

Values must be between 0 and 1. Values of 0.10 or 0.20 are often used for beta. However, 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 1-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.

## Nuisance Parameter Type

Enter the type of nuisance parameter here. Unfortunately, the 2-by-2 table cannot be completely specified by using only the parameters  $P_s$  and  $D_a$  or  $P_s$  and  $R_a$ . One other parameter must be specified. This additional parameter is called a 'nuisance' parameter. It will be assumed to be a known quantity. Several possible choices are available. This option lets you specify which parameter you want to use. In all cases, the value you specify is a proportion.

**P11** is the proportion of subjects that are positive on both tests.

**P00** is the proportion of subjects that are negative on both tests.

**P01** is the proportion of subjects that are negative on the treatment, but positive on the standard.

**P10** is the proportion of subjects that are positive on the treatment, but negative on the standard.

**P11+P00** is the proportion of matches (concordant pairs).

**P01+P10** is the proportion of non-matches (discordant pairs).

**P11/ $P_s$**  is the sensitivity.

## Nuisance Parameter Value

Enter the value of the nuisance parameter that you specified in the 'Nuisance Parameter Type' box. This value is a proportion, so it must be between 0 and 1.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P_1 - P_2$ . P1.0 is the value of P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated. Once P2, D0, and D1 are given, the values of P1.1 and P1.0 can be calculated.

### De (Equivalence Difference)

De is the maximum allowable difference between the standard and treatment proportions that will still result in the conclusion of equivalence. In order to ensure that De is positive, the difference is computed in reverse order. That is,  $D_E = P_S - P_T$ . This parameter is only used when the Test Statistic option is set to 'Difference'.

Only positive values can be entered here. Typical values for this difference are 0.05, 0.10, and 0.20. For two-sided tests, you must have  $|Da| < De$ . For one-sided tests, you must have  $Da > -De$ .

### Da (Actual Difference)

Da is the actual difference between the treatment and standard proportions  $D_A = P_T - P_S$ . Da may be positive, negative, or (usually) zero. This parameter is only used when the Test Statistic option is set to 'Difference'.

For two-sided tests, you must have  $|Da| < De$ . For one-sided tests, you must have  $Da > -De$ .

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P_1 / P_2$ . P1.0 is the value of P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated. Once P2, R0, and R1 are given, the values of P1.0 and P1.1 can be calculated.

### Re (Equivalence Ratio)

Re is the minimum size of the relative risk ratio,  $P_T / P_S$ , that will still result in the conclusion of equivalence. Both equivalence and non-inferiority trials use a value that is less than one. Typical values for this ratio are 0.8 or 0.9.

This parameter is only used when the Test Statistic option is set to 'Ratio'.

### Ra (Actual Ratio)

Enter a value for Ra, the actual relative risk ratio  $P_T / P_S$ . This value is used to generate the value of  $P_T$  using the formula  $P_T = P_S R_a$ . Often this value is set equal to one, but this is not necessary.

This parameter is only used when the Test Statistic option is set to 'Ratio'.



## Example1 - Finding Power

A clinical trial will be conducted to show that a non-invasive MRI test is not inferior to the invasive CTAP reference test. Historical data suggest that the CTAP test is 80% accurate. After careful discussion, the researchers decide that if the MRI test is 75% accurate or better, it will be considered non-inferior. They decide to use a difference test statistic. Thus, the equivalence difference is 0.05. They want to study the power for various sample sizes between 20 and 1000 at the 5% significance level.

They use P01 as the nuisance parameter and look at two values: 0.05 and 0.10.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
De  .....	<b>0.05</b>
Da .....	<b>0.00</b>
Ps .....	<b>0.80</b>
Max N Using Exact Power .....	<b>100</b>
N .....	<b>20 100 200 300 450 600 800 1000</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>Ignored</b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.05 0.10</b>
<b>Axes Tab</b>	
Vertical Range .....	<b>User</b>
Minimum .....	<b>0</b>
Maximum .....	<b>1</b>
Number of Tick Marks .....	<b>10</b>

### Annotated Output

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

## Numeric Results

Numeric Results for a Non-Inferiority (One-Sided) Test of a Difference

Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P01)	Alpha	Beta
0.22938	20	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.77062
0.13717	20	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.86283
0.43625	100	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.56375
0.28895	100	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.71105
0.67771	200	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.32229
0.46318	200	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.53682
0.83244	300	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.16756
0.60369	300	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.39631
0.94287	450	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.05713
0.75745	450	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.24255
0.98206	600	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.01794
0.85657	600	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.14343
0.99651	800	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.00349
0.93172	800	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.06828
0.99937	1000	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.00063
0.96870	1000	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.03130

### Report Definitions

Power is the probability of rejecting a false null hypothesis.

N is the number of subjects, the sample size.

De is the maximum difference between the two proportions that is still called 'equivalent'.

Da is the actual difference between Pt and Ps. That is,  $Da = Pt - Ps$ .

Pt is the response proportion to the treatment (experimental or new) test.

Ps is the response proportion to the standard (reference or old) test.

The Nuisance Parameter is a value that is needed, but is not a direct part of the hypothesis.

Alpha is the probability of rejecting a true null hypothesis.

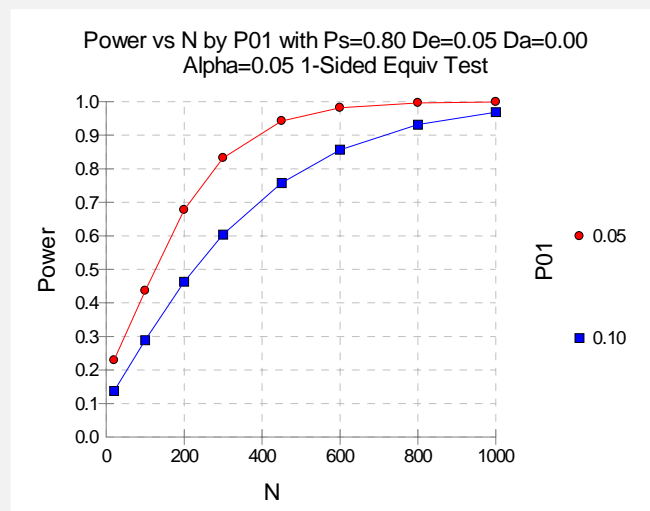
Beta is the probability of accepting a false null hypothesis.

### Summary Statements

A sample size of 20 subjects achieves 23% power at a 5% significance level using a one-sided equivalence test of correlated proportions when the standard proportion is 0.80000, the maximum allowable difference between these proportions that still results in equivalence (the range of equivalence) is 0.05000, and the actual difference of the proportions is 0.00000.

This report shows the power for the indicated scenarios. All of the columns are defined in the 'Report Definitions' section.

## Plot Section



This plot shows the power versus the sample size for the two values of P01. In this example, we see that the value of the nuisance parameter has a large effect on the calculated sample size.

## Example2 - Finding Sample Size

Continuing with Example1, the analysts want to determine the exact sample size necessary to achieve 90% power for both values of the nuisance parameter.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Sample Size (N)</b>
De  .....	<b>0.05</b>
Da .....	<b>0.00</b>
Ps .....	<b>0.80</b>
Max N Using Exact Power .....	<b>100</b>
N .....	<b>Ignored</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.05 0.10</b>

### Annotated Output

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

### Numeric Results

Numeric Results for a Non-Inferiority (One-Sided) Test of a Difference								
Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P01)	Alpha	Beta
0.90026	374	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.09974
0.90014	699	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.09986

This report shows that the sample size required nearly doubles when P01 is changed from 0.05 to 0.10.

## Example3 - Validation using Liu

Liu *et al.* (2002) give an example in which P01 is 0.05, P10 is 0.05, Da is 0.00, De is 0.05, the significance level is 0.025, and the power is 80%. From their Table III, the sample size is 350.

In this example, the value of Ps is arbitrary. We set it at 0.50.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Sample Size (N)</b>
De .....	<b>0.05</b>
Da .....	<b>0.00</b>
Ps.....	<b>0.50</b>
Max N Using Exact Power .....	<b>100</b>
N .....	<b>Ignored</b>
Alpha.....	<b>0.025</b>
Beta.....	<b>0.20</b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.05</b>

### Annotated Output

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

### Numeric Results

Numeric Results for a Non-Inferiority (One-Sided) Test of a Difference								
Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P01)	Alpha	Beta
0.80046	350	0.05000	0.00000	0.50000	0.50000	0.05000	0.02500	0.19954

The calculated sample size of 350 matches the results of Liu et al. (2002).

## Example4 - Validation using Nam and Blackwelder

Nam and Blackwelder (2002) give an example in which  $P_s$  is 0.80,  $P_{10}$  is 0.05,  $R_a$  is 1.00,  $R_e$  is 0.80, the significance level is 0.05, and the power is 80%. From their Table III, the sample size is 34.

Note that their calculations use the approximate formula, so we will set the value of 'Max N Using Exact Power' to '1' so that only the approximate formula is used.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Sample Size (N)</b>
Re.....	<b>0.80</b>
Ra.....	<b>1.00</b>
Ps .....	<b>0.80</b>
Max N Using Exact Power.....	<b>1</b>
N.....	<b>Ignored</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>
Nuisance Parameter Type .....	<b>P10</b>
Nuisance Parameter Value .....	<b>0.05</b>

### Annotated Output

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

### Numeric Results

Numeric Results for a Non-Inferiority (One-Sided) Test of a Difference								
Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P10)	Alpha	Beta
0.80050	34	0.80	1.00	0.80000	0.80000	0.05000	0.05000	0.19950

The calculated sample size of 34 matches the results of Nam and Blackwelder (2002).

## Example5 – Finding Sample Size for a Non-inferiority Test

Researchers have developed a new treatment for migraine headaches which is less expensive than a current standard. The researchers need to show that the proportion of individuals who respond to the new treatment is not inferior to the standard treatment. They want to determine the minimum number of subjects required to achieve 90% power for the test of non-inferiority. The new treatment will be considered non-inferior if its success rate is no less than 90% of the success rate of the standard, which is about 0.65. The sample size required is evaluated for various values (0.3 to 0.9) of the nuisance parameter:  $P_{11}/P_s$  = sensitivity.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Sample Size (N)</b>
Re .....	<b>0.95</b>
Ra .....	<b>1.0</b>
Ps .....	<b>0.65</b>
Max N Using Exact Power .....	<b>1</b>
N .....	<b>Ignored</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
Nuisance Parameter Type .....	<b>P11/Ps (Sensitivity)</b>
Nuisance Parameter Value .....	<b>0.3 to 0.9 by 0.1</b>

### Annotated Output

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

Numeric Results for a Non-Inferiority (One-Sided) Test of a Ratio								
Power	Sample Size (N)	Equiv. Ratio (Re)	Actual Ratio (Ra)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P11/Ps)	Alpha	Beta
	13	0.95	1.00	0.65000	0.65000	0.30000	0.05000	
	13	0.95	1.00	0.65000	0.65000	0.40000	0.05000	
0.90004	5013	0.95	1.00	0.65000	0.65000	0.50000	0.05000	0.09996
0.90001	4013	0.95	1.00	0.65000	0.65000	0.60000	0.05000	0.09999
0.90004	3015	0.95	1.00	0.65000	0.65000	0.70000	0.05000	0.09996
0.90011	2020	0.95	1.00	0.65000	0.65000	0.80000	0.05000	0.09989
0.90016	1035	0.95	1.00	0.65000	0.65000	0.90000	0.05000	0.09984

These scenarios require a large sample size. In fact, the first two rows are blank because the sample size is so large it can't be determined.

## Chapter 165

# Equivalence of Correlated Proportions

The two procedures described in this chapter compute power and sample size for testing equivalence using differences or ratios in designs in which two dichotomous responses are measured on each subject. Each of these options is listed separately on the menus.

When one is interested in showing that two correlated proportions are different, the data are often analyzed with McNemar's test. However, the procedures discussed here are interested in showing equivalence rather than difference. For example, suppose a diagnostic procedure is accurate, but is expensive to apply or has serious side effects. A replacement procedure may be sought which is equally accurate, but is less expensive or has fewer side effects. In this case, we are not interested in showing that the two diagnostic procedures are different, but rather that they are the same. Equivalence tests were designed for this situation.

These tests are often divided into two categories: *equivalence* (two-sided) tests and *non-inferiority* (one-sided) tests. Here, the term *equivalence tests* means that we want to show that two diagnostic procedures are equivalent—that is, their accuracy is about the same. This requires a two-sided hypothesis test. On the other hand, *non-inferiority tests* are used when we want to show that a new (experimental) procedure is no worse than the existing (reference or gold-standard) one. This requires a one-sided hypothesis test.

## Technical Details

The results of a study in which two dichotomous responses are measured on each subject can be displayed in a 2-by-2 table in which one response is shown across the columns and the other is shown down the rows. In the discussion to follow, the columns of the table represent the standard (reference or control) response and the rows represent the treatment (experimental) response. The outcome probabilities can be classified into the following table.

Experimental Diagnosis	Standard Yes	Diagnosis No	Total
Yes	$p_{11}$	$p_{10}$	$P_T$
No	$p_{01}$	$p_{00}$	$1 - P_T$
Total	$P_S$	$1 - P_S$	1

In this table,  $p_{ij} = p_{Treatment, Standard}$ . That is, the first subscript represents the response of the new, experimental procedure while the second subscript represents the response of the standard procedure. Thus,  $p_{01}$  represents the proportion having a negative treatment response and a positive standard response.

To aid in interpretation, analysts have developed a few proportions that summarize the table. Three of the most popular ratios are *sensitivity*, *specificity*, and *prevalence*.

### Sensitivity

Sensitivity is the proportion of subjects with a positive standard response who also have a positive experimental response. In terms of proportions from the 2-by-2 table,

$$\text{Sensitivity} = p_{11} / (p_{01} + p_{11}) = p_{11} / P_S$$

### Specificity

Specificity is the proportion of subjects with a negative standard response who also have a negative experimental response. In terms of proportions from the 2-by-2 table,

$$\text{Specificity} = p_{00} / (p_{10} + p_{00})$$

### Prevalence

Prevalence is the overall proportion of individuals with the disease (or feature of interest). In terms of proportions from the 2-by-2 table,

$$\text{Prevalence} = P_S$$



The outcome counts from a sample of  $n$  subjects can be classified into the following table.

Experimental Diagnosis	Standard Yes	Diagnosis No	Total
Yes	$n_{11}$	$n_{10}$	$n_T$
No	$n_{01}$	$n_{00}$	$n - n_T$
Total	$n_S$	$n - n_S$	$n$

Note that  $n_{11} + n_{00}$  is the number of matches (*concordant pairs*) and  $n_{10} + n_{01}$  is the number of *discordant pairs*.

The hypothesis of interest concerns the two marginal probabilities  $P_T$  and  $P_S$ .  $P_S$  represents the accuracy or success of the standard test and  $P_T$  represents the accuracy or success of the new, experimental test. Equivalence is defined in terms of either the difference,  $D = P_T - P_S$ , or the relative risk ratio,  $R = P_T / P_S$ , of these two proportions. The choice between  $D$  and  $R$  will usually lead to different sample sizes to achieve the same power.

## Equivalence Hypotheses using Difference

This section is based on Liu, Hsueh, Hsieh and Chen (2002). We refer you to that paper for complete details.

The hypotheses of equivalence in terms of the difference are

$$H_0: P_T - P_S \geq D_E \quad \text{or} \quad H_0: P_T - P_S \leq -D_E \quad \text{versus} \quad H_1: -D_E < P_T - P_S < D_E$$

These hypotheses can be decomposed into two sets of one-sided hypotheses

$$H_{0L}: P_T - P_S \leq -D_E \quad \text{versus} \quad H_{1L}: P_T - P_S > -D_E$$

and

$$H_{0U}: P_T - P_S \geq D_E \quad \text{versus} \quad H_{1U}: P_T - P_S < D_E$$

The hypothesis test of equivalence with type I error rate  $\alpha$  is conducted by computing a  $100(1 - 2\alpha)\%$  confidence interval for  $P_T - P_S$  and determining if this interval is wholly contained between  $-D_E$  and  $D_E$ . This confidence interval approach is often recommended by regulatory agencies.

Liu et al. (2002) discuss the RMLE-based (score) method for constructing these confidence intervals. This method is based on (developed by, described by) Nam (1997).

## Asymptotic Tests

An asymptotic test for testing  $H_{0L}$  versus  $H_{1L}$  is given by

$$Z_L = \frac{\hat{D} + D_E}{\hat{\sigma}} = \frac{c + nD_E}{\sqrt{d - n\hat{D}^2}} \geq z_\alpha$$

where

$$\hat{D} = \frac{n_T}{n} - \frac{n_S}{n} = \frac{n_{10}}{n} - \frac{n_{01}}{n}$$

$$d = n_{10} + n_{01}$$

$$c = n_{10} - n_{01}$$

and  $z_\alpha$  is the standard normal deviate having  $\alpha$  in the right tail.

Similarly, an asymptotic test for testing  $H_{0U}$  versus  $H_{1U}$  is given by

$$Z_U = \frac{\hat{D} - D_E}{\hat{\sigma}} = \frac{c - nD_E}{\sqrt{d - n\hat{D}^2}} \leq -z_\alpha.$$

Equivalence is concluded if both the tests on  $Z_L$  and  $Z_U$  are rejected..

An estimate of  $\hat{\sigma}$  based on the RMLE-based (score) procedure of Nam (1997) uses the estimates

$$\tilde{\sigma}_L = \sqrt{\frac{\tilde{p}_{L,10} + \tilde{p}_{L,01} - D_E^2}{n}}$$

and

$$\tilde{\sigma}_U = \sqrt{\frac{\tilde{p}_{U,10} + \tilde{p}_{U,01} - D_E^2}{n}}$$

where

$$\tilde{p}_{L,10} = \frac{-\tilde{a}_L + \sqrt{\tilde{a}_L^2 - 8\tilde{b}_L}}{4}$$

$$\tilde{p}_{L,01} = \tilde{p}_{L,10} - D_E$$

$$\tilde{p}_{U,10} = \frac{-\tilde{a}_U + \sqrt{\tilde{a}_U^2 - 8\tilde{b}_U}}{4}$$

$$\tilde{p}_{U,01} = \tilde{p}_{U,10} + D_E$$

$$\tilde{a}_{L,01} = -\hat{D}(1 - D_E) - 2(\hat{p}_{01} + D_E)$$

$$\tilde{b}_{L,01} = D_E(1 + D_E)\hat{p}_{01}$$

$$\tilde{a}_{U,01} = -\hat{D}(1 + D_E) - 2(\hat{p}_{01} - D_E)$$

$$\tilde{b}_{U,01} = -D_E(1 - D_E)\hat{p}_{01}$$

Note that the ICH E9 guideline (see Lewis (1999)) suggests using a significance level of  $\alpha / 2$  when testing this hypothesis.

## Power Formula

The power when the actual difference is  $D_A$  can be evaluated exactly using the multinomial distribution. However, when the sample size is above a user-set level, we use a normal approximation to this distribution which leads to

$$1 - \beta(D_A) = \begin{cases} \Phi(c_U) - \Phi(c_L) & \text{if } c_U - c_L > 0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$c_U = \frac{-D_A}{\sigma} + \frac{D_E}{\sigma} - \frac{z_\alpha}{w_U}$$

$$c_L = \frac{-D_A}{\sigma} - \frac{D_E}{\sigma} + \frac{z_\alpha}{w_L}$$

$$\sigma = \sqrt{\frac{p_{01} + p_{10} - D_A}{n}}$$

$$w_L = \sqrt{\frac{2p_{01} + D_A - D_A^2}{2\bar{p}_{L,01} - D_E - D_E^2}}$$

$$w_U = \sqrt{\frac{2p_{01} + D_A - D_A^2}{2\bar{p}_{U,01} + D_E - D_E^2}}$$

$$\bar{p}_{L,01} = \frac{-a_L + \sqrt{a_L^2 - 8b_L}}{4}$$

$$\bar{p}_{U,01} = \frac{-a_U + \sqrt{a_U^2 - 8b_U}}{4}$$

$$a_L = -D_A(1 - D_E) - 2(p_{01} + D_E)$$

$$b_L = D_E(1 + D_E)p_{01}$$

$$a_U = -D_A(1 + D_E) - 2(p_{01} - D_E)$$

$$b_U = -D_E(1 - D_E)p_{01}$$

## Equivalence Hypotheses using Ratios

For the two-sided (equivalence) case when  $R_E < 1$ , the statistical hypotheses are

$$H_0: P_T / P_S \leq 1 / R_E \text{ or } P_T / P_S \geq R_E \quad \text{versus} \quad H_1: R_E < P_T / P_S < 1 / R_E$$

These can be decomposed into two sets of one-sided hypotheses

$$H_{0L}: P_T / P_S \leq R_E \quad \text{versus} \quad H_{1L}: P_T / P_S > R_E$$

and

$$H_{0U}: P_T / P_S \geq 1 / R_E \quad \text{versus} \quad H_{1U}: P_T / P_S < 1 / R_E$$

Note that the first set of one-sided hypotheses,  $H_{0L}$  versus  $H_{1L}$ , is referred to as the hypotheses of non-inferiority.

The following is based on Nam and Blackwelder (2002). We refer you to this paper for the complete details of which we will only provide a brief summary here.

## Test Statistics

The test statistic for an asymptotic test based on constrained maximum likelihood for large  $n$  is given by

$$Z(R_E) = \sqrt{\frac{n(\hat{P}_T - R_E \hat{P}_S)}{R_E(\tilde{p}_{10} + \tilde{p}_{01})}}$$

where

$$\tilde{p}_{10} = \frac{-\hat{P}_T + R_E^2(\hat{P}_S + 2\hat{p}_{10}) + \sqrt{(\hat{P}_T - R_E^2 \hat{P}_S)^2 + 4R_E^2 \hat{p}_{10} \hat{p}_{01}}}{2R_E(R_E + 1)}$$

$$\tilde{p}_{01} = R_E \tilde{p}_{10} - (R_E - 1)(1 - \hat{p}_{00})$$

$$\hat{p}_{01} = \frac{n_{01}}{n}, \hat{p}_{10} = \frac{n_{10}}{n}, \hat{P}_T = \frac{n_{10} + n_{11}}{n}, \hat{P}_S = \frac{n_{01} + n_{11}}{n}$$

Note that the above applies to a one-sided test. When using a two-sided test, we calculate both  $Z(R_E)$  and  $Z(1/R_E)$  using the above formula.

## Power Formula

The power of the one-sided procedure when the true value of the relative risk ratio is  $R_E$  can be evaluated exactly using the multinomial distribution. When  $n$  is large, we use a normal approximation to the multinomial distribution which leads to

$$\beta(R_A) = \Phi(c_U)$$

where

$$c_U = \frac{z_{1-\alpha} \sqrt{\bar{V}_0(T_0)} - E_1(T_0)}{\sqrt{V_1(T_0)}}$$

$$\bar{V}_0(T_0) = \frac{R_E(\bar{p}_{10} + \bar{p}_{01})}{n}$$

$$E_1(T_0) = (R_A - R_E)P_S$$

$$V_1(T_0) = \frac{(R_A + R_E^2)P_S - 2R_E p_{11} - (R_A - R_E)^2 P_S^2}{n}$$

$$\bar{p}_{10} = \frac{-P_T + R_E^2(P_S + 2p_{10}) + \sqrt{(P_T - R_E^2 P_S)^2 + 4R_E^2 p_{10} p_{01}}}{2R_E(R_E + 1)}$$

$$\bar{p}_{01} = R_E \bar{p}_{10} - (R_E - 1)(1 - p_{00})$$

## Nuisance Parameter

Unfortunately, the 2-by-2 table includes four parameters  $p_{11}$ ,  $p_{10}$ ,  $p_{01}$ , and  $p_{00}$ , but the power specifications above only specify two parameters:  $P_S$  and  $D_A$  or  $R_A$ . A third parameter is defined implicitly since the sum of the four parameters is one. One parameter, known as a nuisance parameter, remains unaccounted for. This parameter must be addressed to fully specify the problem. This fourth parameter can be specified by specifying any one of the following:

$p_{11}$ ,  $p_{10}$ ,  $p_{01}$ ,  $p_{00}$ ,  $p_{10} + p_{01}$ ,  $p_{11} + p_{00}$ , or the sensitivity of the experimental response,  $p_{11} / P_S$ .

It may be difficult to specify a reasonable value for the nuisance parameter since its value may not be even approximately known until after the study is conducted. Because of this, we suggest that you calculate power or sample size for a range of values of the nuisance parameter. This will allow you to determine how sensitive the results are to its value.

## 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 at the beginning of this manual.

## Data Tab (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers two procedures which have different options. This section documents options that are common to both procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Beta and Power* or *N*.

### Ps (Standard Proportion)

This is the proportion of yes's (or successes),  $P_s$ , when subjects received the standard treatment. This value or a good estimate is often available from previous studies.

Note that this value does not matter when the Nuisance Parameter Type is set to 'P01' (or 'P10'), as long as it is greater than P01 (or P10).

You may enter a set of values separated by blank spaces. For example, you could enter *0.50 0.60 0.70*. Values between, but not including, 0 and 1 are permitted.

### Max N Using Exact Power

Specify the maximum value of N (sample size) for which you would like an exact power calculation based on the multinomial distribution. Sample Sizes greater than this value will use the asymptotic approximation given in the documentation. The exact calculation of the multinomial distribution becomes very time consuming for  $N > 200$ . For most cases, when  $N > 200$ , the difference between the exact and approximate calculations is small. For  $N > 200$ , the length of time needed to calculate the exact answer may become prohibitive. However, as the speed of computers increases, it will become faster and easier to calculate the exact power for larger values of N.

If you want all calculations to use exact results, enter '1000' here.

If you want all calculations to use the quick approximations, enter '1' here.

### N (Sample Size)

Enter a value for the sample size. This value must be greater than two. You may enter a range of values such as *10 to 100 by 10*.

### 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 conclude equivalence or inferiority.

## 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 conclude equivalence or non-inferiority when in fact it is true.

Values must be between 0 and 1. Values of 0.10 or 0.20 are often used for beta. However, 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 1-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.

## Nuisance Parameter Type

Enter the type of nuisance parameter here. Unfortunately, the 2-by-2 table cannot be completely specified by using only the parameters  $P_s$  and  $D_a$  or  $P_s$  and  $R_a$ . One other parameter must be specified. This additional parameter is called a 'nuisance' parameter. It will be assumed to be a known quantity. Several possible choices are available. This option lets you specify which parameter you want to use. In all cases, the value you specify is a proportion.

**P11** is the proportion of subjects that are positive on both tests.

**P00** is the proportion of subjects that are negative on both tests.

**P01** is the proportion of subjects that are negative on the treatment, but positive on the standard.

**P10** is the proportion of subjects that are positive on the treatment, but negative on the standard.

**P11+P00** is the proportion of matches (concordant pairs).

**P01+P10** is the proportion of non-matches (discordant pairs).

**P11/Ps** is the sensitivity.

## Nuisance Parameter Value

Enter the value of the nuisance parameter that you specified in the 'Nuisance Parameter Type' box. This value is a proportion, so it must be between 0 and 1.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P_1 - P_2$ . P1.0 is the value of P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated. Once P2, D0, and D1 are given, the values of P1.1 and P1.0 can be calculated.

### De (Equivalence Difference)

De is the maximum allowable difference between the standard and treatment proportions that will still result in the conclusion of equivalence. In order to ensure that De is positive, the difference is computed in reverse order. That is,  $D_E = P_S - P_T$ . This parameter is only used when the Test Statistic option is set to *Difference*.

Only positive values can be entered here. Typical values for this difference are 0.05, 0.10, and 0.20. For two-sided tests, you must have  $|Da| < De$ . For one-sided tests, you must have  $Da > -De$ .

### Da (Actual Difference)

Da is the actual difference between the treatment and standard proportions  $D_A = P_T - P_S$ . Da may be positive, negative, or (usually) zero. This parameter is only used when the Test Statistic option is set to *Difference*.

For two-sided tests, you must have  $|Da| < De$ . For one-sided tests, you must have  $Da > -De$ .

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P_1 / P_2$ . P1.0 is the value of P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated. Once P2, R0, and R1 are given, the values of P1.0 and P1.1 can be calculated.

### Re (Equivalence Ratio)

Re is the minimum size of the relative risk ratio,  $P_T / P_S$ , that will still result in the conclusion of equivalence. Both equivalence and non-inferiority trials use a value that is less than one. Typical values for this ratio are 0.8 or 0.9.

This parameter is only used when the Test Statistic option is set to *Ratio*.

### Ra (Actual Ratio)

Enter a value for Ra, the actual relative risk ratio  $P_T / P_S$ . This value is used to generate the value of  $P_T$  using the formula  $P_T = P_S R_a$ . Often this value is set equal to one, but this is not necessary.

This parameter is only used when the Test Statistic option is set to *Ratio*.



## Example1 - Finding Power

A clinical trial will be conducted to show that a non-invasive MRI test is equivalent to the invasive CTAP reference test. Historical data suggest that the CTAP test is 80% accurate. After careful discussion, the researchers decide that if the MRI test is five percentage points of the CTAP, it will be considered equivalent. They decide to use a difference test statistic. Thus, the equivalence difference is 0.05. They want to study the power for various sample sizes between 20 and 1000 at the 5% significance level. They decide to use the approximate power calculations, so they set the 'Max N Using Exact Power' option to 2.

They use P01 as the nuisance parameter and look at two values: 0.05 and 0.10.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
De  .....	<b>0.05</b>
Da .....	<b>0.00</b>
Ps .....	<b>0.80</b>
Max N Using Exact Power .....	<b>2</b>
N .....	<b>20 100 200 300 450 600 800 1000</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>Ignored</b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.05 0.10</b>
<b>Axes Tab</b>	
Vertical Range .....	<b>User</b>
Minimum .....	<b>0</b>
Maximum .....	<b>1</b>
Number of Tick Marks .....	<b>10</b>

## Annotated Output

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

### Numeric Results

#### Numeric Results for an Equivalence (Two-Sided) Test of a Difference

Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P01)	Alpha	Beta
0.00000	20	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	1.00000
0.00000	20	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	1.00000
0.00000	100	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	1.00000
0.00000	100	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	1.00000
0.35542	200	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.64458
0.00000	200	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	1.00000
0.66488	300	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.33512
0.20739	300	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.79261
0.88574	450	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.11426
0.51491	450	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.48509
0.96411	600	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.03589
0.71314	600	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.28686
0.99301	800	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.00699
0.86344	800	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.13656
0.99874	1000	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.00126
0.93739	1000	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.06261

#### Report Definitions

Power is the probability of rejecting a false null hypothesis.

N is the number of subjects, the sample size.

De is the maximum difference between the two proportions that is still called 'equivalent'.

Da is the actual difference between Pt and Ps. That is,  $Da = Pt - Ps$ .

Pt is the response proportion to the treatment (experimental or new) test.

Ps is the response proportion to the standard (reference or old) test.

The Nuisance Parameter is a value that is needed, but is not a direct part of the hypothesis.

Alpha is the probability of rejecting a true null hypothesis.

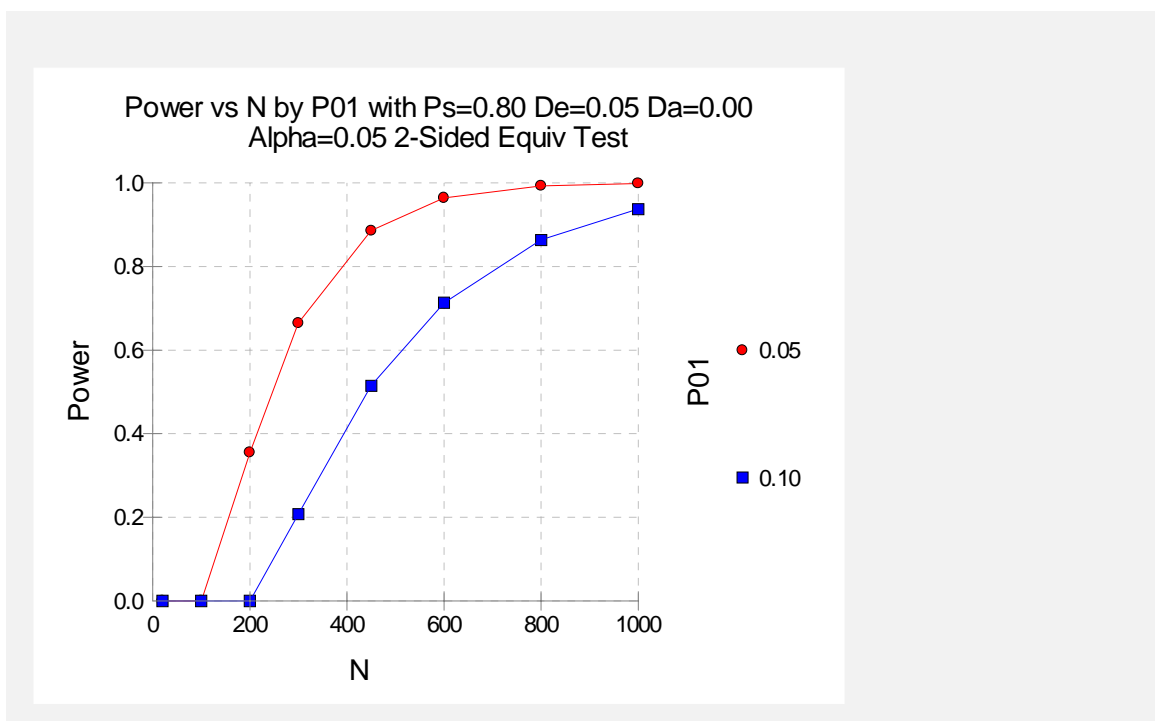
Beta is the probability of accepting a false null hypothesis.

#### Summary Statements

A sample size of 20 subjects achieves 0% power at a 5% significance level using a two-sided equivalence test of correlated proportions when the standard proportion is 0.80000, the maximum allowable difference between these proportions that still results in equivalence (the range of equivalence) is 0.05000, and the actual difference of the proportions is 0.00000.

This report shows the power for the indicated scenarios. All of the columns are defined in the 'Report Definitions' section.

## Plot Section



This plot shows the power versus the sample size for the two values of P01. In this example, we see that the value of the nuisance parameter has a large effect on the calculated sample size.

## Example2 - Finding Sample Size

Continuing with Example1, the analysts want to determine the exact sample size necessary to achieve 90% power for both values of the nuisance parameter.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Sample Size (N)</b>
De .....	<b>0.05</b>
Da .....	<b>0.00</b>
Ps.....	<b>0.80</b>
Max N Using Exact Power .....	<b>2</b>
N .....	<b>Ignored</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.10</b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.05 0.10</b>

### Annotated Output

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

### Numeric Results

Numeric Results for an Equivalence (Two-Sided) Test of a Difference								
Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P01)	Alpha	Beta
0.90019	468	0.05000	0.00000	0.80000	0.80000	0.05000	0.05000	0.09981
0.90002	881	0.05000	0.00000	0.80000	0.80000	0.10000	0.05000	0.09998

This report shows that the sample size required nearly doubles when P01 is changed from 0.05 to 0.10.

## Example3 - Validation using Liu

Liu *et al.* (2002) page 238 give a table of power values for sample sizes of 50, 100, and 200 when the significance level is 0.05. From this table, we find that when  $P_{01}$  is 0.10,  $P_{10}$  is 0.10,  $D_a = P_{01} - P_{10} = 0.00$ , and  $D_e$  is 0.10, and the three power values are 0.026, 0.417, and 0.861 for the column head 'RMLE-based Without CC' (this is the case we use).

In their calculations, they round the  $z$  value to 1.64. This corresponds to an alpha value of 0.0505025835. So that our results match, we will use this value for alpha rather than 0.05.

In this example, the value of  $P_s$  is not used. We set it at 0.50. Also, we set the Max N value to 200 so that the exact values will be calculated.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
De  .....	<b>0.1</b>
Da .....	<b>0.0</b>
Ps .....	<b>0.5</b>
Max N Using Exact Power .....	<b>200</b>
N .....	<b>50 100 200</b>
Alpha .....	<b>0.0505025835</b>
Beta .....	<b><i>Ignored since this is the Find setting</i></b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.1</b>

### Annotated Output

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

### Numeric Results

Numeric Results for an Equivalence (Two-Sided) Test of a Difference								
Power	Sample Size (N)	Equiv. Difference (De)	Actual Difference (Da)	Treatment Proportion (Pt)	Standard Proportion (Ps)	Nuisance Parameter (P01)	Alpha	Beta
0.02614	50	0.10000	0.00000	0.50000	0.50000	0.10000	0.05050	0.97386
0.41741	100	0.10000	0.00000	0.50000	0.50000	0.10000	0.05050	0.58259
0.86080	200	0.10000	0.00000	0.50000	0.50000	0.10000	0.05050	0.13920

As you can see, the values computed by *PASS* match the results of Liu et al. (2002).

## Example4 – Finding Power Following an Experiment

An experiment involving a single group of 57 subjects was run to show that a new treatment was equivalent to a previously used standard. Historically, the standard treatment has had a 48% success rate. The new treatment is known to have similar side effects to the standard, but is much less expensive. The treatments were to be considered equivalent if the success rate of the new treatment is within 10% of the success rate of the standard.

To compare the new and standard treatments, each of the 57 subjects received both treatments with a washout period between them. Thus, the proportions based on the two treatments are correlated. Of the 57 subjects, 18 responded to both treatments, 20 did not respond to either treatment, 9 responded to the new treatment but not the standard, and 10 responded to the standard but not the new treatment. The proportion responding to the new treatment is  $(18+9)/57 = 0.4737$ . The proportion responding to the standard is  $(18+10)/57 = 0.4912$ . The difference is 0.0175, lower than the threshold for equivalence, but the resulting p-value was 0.3358, indicating the two treatments could not be deemed equivalent at the 0.05 level. Note that McNemar's test only uses the discordant pairs, so the effective size of this study is really only  $9 + 10 = 19$ , although 57 subjects were investigated. The researchers want to know the power of the test they used.

It may be the inclination of the researchers to use the observed difference in proportions for calculating power. The p-value, however, is based on the maximum allowable difference for equivalence, which is 10% of 0.48, or 0.048. This is the number that should be used in the power calculation. The experiment gave a value of P01 of  $10/28 = 0.36$ . The power of the experiment is near zero for all values of P01 less than 0.10. We calculate the power for a variety of nuisance parameter values ( $P01 = 0.01, 0.03, 0.05$ , and  $0.10$ ) to monitor its effect. Because it is in fact believed that the success rates are equivalent for the two treatments, the specified actual difference is set to 0.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
De .....	<b>0.048</b>
Da .....	<b>0.00</b>
Ps .....	<b>0.48</b>
Max N Using Exact Power .....	<b>200</b>
N .....	<b>57</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>Ignored</b>
Nuisance Parameter Type .....	<b>P01</b>
Nuisance Parameter Value .....	<b>0.01 0.03 0.05 0.10</b>

## Annotated Output

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

### Numeric Results

**Numeric Results for an Equivalence (Two-Sided) Test of a Difference**

<b>Power</b>	<b>Sample Size (N)</b>	<b>Equiv. Difference (De)</b>	<b>Actual Difference (Da)</b>	<b>Treatment Proportion (Pt)</b>	<b>Standard Proportion (Ps)</b>	<b>Nuisance Parameter (P01)</b>	<b>Alpha</b>	<b>Beta</b>
0.31614	57	0.04800	0.00000	0.48000	0.48000	0.01000	0.05000	0.68386
0.02940	57	0.04800	0.00000	0.48000	0.48000	0.03000	0.05000	0.97060
0.00247	57	0.04800	0.00000	0.48000	0.48000	0.05000	0.05000	0.99753
0.00000	57	0.04800	0.00000	0.48000	0.48000	0.10000	0.05000	1.00000

Note that there is no power for value of P01 greater than 0.05. This is probably due to the low number of discordant pairs.





## Chapter 200

# Two Independent Proportions – Inequality

## Introduction

This module computes power and sample size for hypothesis tests of the difference, ratio, or odds ratio of two independent proportions. The term *null case* in the chapter title refers to the value of the null hypothesis. The test statistics analyzed by this procedure assume that the difference between the two proportions is zero or their ratio is one under the null hypothesis. The *non-null (offset) case* is discussed in another procedure. This procedure computes and compares the power achieved by each of several test statistics that have been proposed.

For example, suppose you want to compare two methods for treating cancer. Your experimental design might be as follows. Select a sample of patients and randomly assign half to one method and half to the other. After five years, determine the proportion surviving in each group and test whether the difference in the proportions is significantly different from zero.

The power calculations assume that random samples are drawn from two separate populations.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Technical Details

Suppose you have two populations from which dichotomous (binary) responses will be recorded. The probability (or risk) of obtaining the event of interest in population 1 (the treatment group) is  $p_1$  and in population 2 (the control group) is  $p_2$ . The corresponding failure proportions are given by  $q_1 = 1 - p_1$  and  $q_2 = 1 - p_2$ .

The assumption is made that the responses from each group follow a binomial distribution. This means that the event probability,  $p_i$ , is the same for all subjects within the group and that the response from one subject is independent of that of any other subject.

Random samples of  $m$  and  $n$  individuals are obtained from these two populations. The data from these samples can be displayed in a 2-by-2 contingency table as follows

Group	Success	Failure	Total
Treatment	$a$	$c$	$m$
Control	$b$	$d$	$n$
Total	$s$	$f$	$N$

The following alternative notation is also used.

Group	Success	Failure	Total
Treatment	$x_{11}$	$x_{12}$	$n_1$
Control	$x_{21}$	$x_{22}$	$n_2$
Total	$m_1$	$m_2$	$N$

The binomial proportions  $p_1$  and  $p_2$  are estimated from these data using the formulae

$$\hat{p}_1 = \frac{a}{m} = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{b}{n} = \frac{x_{21}}{n_2}$$

## Comparing Two Proportions

When analyzing studies such as this, one usually wants to compare the two binomial probabilities,  $p_1$  and  $p_2$ . Common measures for comparing these quantities are the difference and the ratio. If the binomial probabilities are expressed in terms of odds rather than probabilities, another common measure is the odds ratio. Mathematically, these comparison parameters are

Parameter	Computation
Difference	$\delta = p_1 - p_2$
Risk Ratio	$\phi = p_1 / p_2$
Odds Ratio	$\psi = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 q_2}{p_2 q_1}$

The tests analyzed by this routine are for the *null case*. This refers to the values of the above parameters under the null hypothesis. In the *null case*, the difference is zero and the ratios are one under the null hypothesis. In the *non-null case*, discussed in another chapter, the difference is some value other than zero and the ratios are some value other than one. The non-null case often appears in equivalence and non-inferiority testing.

# Hypothesis Tests

Several statistical tests have been developed for testing the inequality of two proportions. For large samples, the powers of the various tests are about the same. However, for small samples, the differences in the powers can be quite large. Hence, it is important to base the power analysis on the test statistic that will be used to analyze the data. If you have not selected a test statistic, you may wish to determine which one offers the best power in your situation. No single test is the champion in every situation, so you must compare the powers of the various tests to determine which to use.

## Difference

The (risk) difference,  $\delta = p_1 - p_2$ , is perhaps the most direct measure for comparing two proportions. Three sets of statistical hypotheses can be formulated:

1.  $H_0: p_1 - p_2 = 0$  versus  $H_1: p_1 - p_2 \neq 0$ ; this is often called the *two-tailed test*.
2.  $H_0: p_1 - p_2 \leq 0$  versus  $H_1: p_1 - p_2 > 0$ ; this is often called the *upper-tailed test*.
3.  $H_0: p_1 - p_2 \geq 0$  versus  $H_1: p_1 - p_2 < 0$ ; this is often called the *lower-tailed test*.

The traditional approach for testing these hypotheses has been to use the Pearson chi-square test for large samples, the Yates chi-square for intermediate sample sizes, and the Fisher Exact test for small samples. Recently, some authors have begun questioning this solution. For example, based on exact enumeration, Upton (1982) and D'Agostino (1988) conclude that the Fisher Exact test and Yates test should never be used.

## Ratio

The (risk) ratio,  $\phi = p_1 / p_2$ , is often preferred to the difference when the baseline proportion is small (less than 0.1) or large (greater than 0.9) because it expresses the difference as a percentage rather than an amount. In this null case, the null hypothesized ratio of proportions,  $\phi_0$ , is one.

Three sets of statistical hypotheses can be formulated:

1.  $H_0: p_1 / p_2 = \phi_0$  versus  $H_1: p_1 / p_2 \neq \phi_0$ ; this is often called the *two-tailed test*.
2.  $H_0: p_1 / p_2 \leq \phi_0$  versus  $H_1: p_1 / p_2 > \phi_0$ ; this is often called the *upper-tailed test*.
3.  $H_0: p_1 / p_2 \geq \phi_0$  versus  $H_1: p_1 / p_2 < \phi_0$ ; this is often called the *lower-tailed test*.

## Odds Ratio

The odds ratio,  $\psi = \frac{o_1}{o_2} = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 q_2}{p_2 q_1}$ , is sometimes used to compare the two

proportions because of its statistical properties and because some experimental designs require its use. In this null case, the null hypothesized odds ratio,  $\psi_0$ , is one. Three sets of statistical hypotheses can be formulated:

1.  $H_0: \psi = \psi_0$  versus  $H_1: \psi \neq \psi_0$ ; this is often called the *two-tailed test*.
2.  $H_0: \psi \leq \psi_0$  versus  $H_1: \psi > \psi_0$ ; this is often called the *upper-tailed test*.
3.  $H_0: \psi \geq \psi_0$  versus  $H_1: \psi < \psi_0$ ; this is often called the *lower-tailed test*.

## Power Calculation

The power for a test statistic that is based on the normal approximation can be computed exactly using two binomial distributions. The following steps are taken to compute the power of such a test.

1. Find the critical value (or values in the case of a two-sided test) using the standard normal distribution. The critical value,  $z_{critical}$ , is that value of  $z$  that leaves exactly the target value of alpha in the appropriate tail of the normal distribution. For example, for an upper-tailed test with a target alpha of 0.05, the critical value is 1.645.
2. Compute the value of the test statistic,  $z_t$ , for every combination of  $x_{11}$  and  $x_{21}$ . Note that  $x_{11}$  ranges from 0 to  $n_1$ , and  $x_{21}$  ranges from 0 to  $n_2$ . A small value (around 0.0001) can be added to the zero cell counts to avoid numerical problems that occur when the cell value is zero.
3. If  $z_t > z_{critical}$ , the combination is in the rejection region. Call all combinations of  $x_{11}$  and  $x_{21}$  that lead to a rejection the set  $A$ .
4. Compute the power for given values of  $p_1$  and  $p_2$  as

$$1 - \beta = \sum_A \binom{n_1}{x_{11}} p_1^{x_{11}} q_1^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_2^{x_{21}} q_2^{n_2 - x_{21}}$$

5. Compute the actual value of alpha achieved by the design by substituting  $p_2$  for  $p_1$

$$\alpha^* = \sum_A \binom{n_1}{x_{11}} p_1^{x_{11}} q_1^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_1^{x_{21}} q_1^{n_2 - x_{21}}$$

When the values of  $n_1$  and  $n_2$  are large (say over 200), these formulas may take a little time to evaluate. In this case, a large sample approximation may be used.

## Test Statistics

The various test statistics that are available in this routine are listed next.

### Fisher's Exact Test

The most useful reference we found for power analysis of Fisher's Exact test was in the StatXact 5 (2001) documentation. The material present here is summarized from Section 26.3 (pages 866 – 870) of the StatXact-5 documentation. In this case, the test statistic is

$$T = -\ln \left[ \frac{\binom{n_1}{x_1} \binom{n_2}{x_2}}{\binom{N}{m}} \right]$$

The null distribution of T is based on the hypergeometric distribution. It is given by

$$\Pr(T \geq t | m, H_0) = \sum_{A(m)} \left[ \frac{\binom{n_1}{x_1} \binom{n_2}{x_2}}{\binom{N}{m}} \right]$$

where

$$A(m) = \{ \text{all pairs } x_1, x_2 \text{ such that } x_1 + x_2 = m, \text{ given } T \geq t \}$$

Conditional on  $m$ , the critical value,  $t_\alpha$ , is the smallest value of  $t$  such that

$$\Pr(T \geq t_\alpha | m, H_0) \leq \alpha$$

The power is defined as

$$1 - \beta = \sum_{m=0}^N P(m) \Pr(T \geq t_\alpha | m, H_1)$$

where

$$\begin{aligned} \Pr(T \geq t_\alpha | m, H_1) &= \sum_{A(m, T \geq t_\alpha)} \left[ \frac{b(x_1, n_1, p_1) b(x_2, n_2, p_2)}{\sum_{A(m)} b(x_1, n_1, p_1) b(x_2, n_2, p_2)} \right] \\ P(m) &= \Pr(x_1 + x_2 = m | H_1) \\ &= b(x_1, n_1, p_1) b(x_2, n_2, p_2) \\ b(x, n, p) &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

When either group's sample size is greater than the Maximum N1 or N2 limit, an approximation is used based on the pooled, continuity corrected Chi-Square test.

## Chi-Square Test (Pooled and Unpooled)

This test statistic was first proposed by Karl Pearson in 1900. Although this test is usually expressed directly as a Chi-Square statistic, it is expressed here as a  $z$  statistic so that it can be more easily used for one-sided hypothesis testing.

Both *pooled* and *unpooled* versions of this test have been discussed in the statistical literature. The pooling refers to the way in which the standard error is estimated. In the pooled version, the two proportions are averaged, and only one proportion is used to estimate the standard error. In the unpooled version, the two proportions are used separately.

The formula for the test statistic is

$$z_t = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_D}$$

### Pooled Version

$$\hat{\sigma}_D = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

### Unpooled Version

$$\hat{\sigma}_D = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

## Power

The power of this test is computed using the enumeration procedure described above. For large sample sizes, the following approximation is used.

1. Find the critical value (or values in the case of a two-sided test) using the standard normal distribution. The critical value is that value of  $z$  that leaves exactly the target value of alpha in the tail.
2. Use the normal approximation to binomial distribution to compute binomial probabilities, Compute the power using

$$1 - \beta = \Pr\left(Z < z_\alpha \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} + \frac{p_1 - p_2}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}\right)$$

where

$$\bar{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \text{ and } \bar{q} = 1 - \bar{p}$$

## Chi-Square Test with Continuity Correction

Frank Yates is credited with proposing a correction to the Pearson Chi-Square test for the lack of continuity in the binomial distribution. However, the correction was in common use when he proposed it in 1922.

Both *pooled* and *unpooled* versions of this test have been discussed in the statistical literature. The pooling refers to the way in which the standard error is estimated. In the pooled version, the two proportions are averaged, and only one proportion is used to estimate the standard error. In the unpooled version, the two proportions are used separately.

The continuity corrected z-test is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) + \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_D}$$

where  $F$  is -1 for lower-tailed, 1 for upper-tailed, and both -1 and 1 for two-sided hypotheses.

### Pooled Version

$$\hat{\sigma}_D = \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

### Unpooled Version

$$\hat{\sigma}_D = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

### Power

The power of this test is computed using the enumeration procedure described for the z-test above. For large samples, approximate results based on the normal approximation to the binomial are used.

## Conditional Mantel Haenszel Test

The conditional Mantel Haenszel test, see Lachin (2000) page 40, is based on the *index frequency*,  $x_{11}$ , from the 2x2 table. The formula for the z-statistic is

$$z = \frac{x_{11} - E(x_{11})}{\sqrt{V_c(x_{11})}}$$

where

$$E(x_{11}) = \frac{n_1 m_1}{N}$$

$$V_c(x_{11}) = \frac{n_1 n_2 m_1 m_2}{N^2 (N - 1)}$$

### Power

The power of this test is computed using the enumeration procedure described above.

## Likelihood Ratio Test

In 1935, Wilks showed that the following quantity has a chi-square distribution with one degree of freedom. Using this test statistic to compare proportions is presented, among other places, in Upton (1982). The likelihood ratio test statistic is computed as

$$LR = 2 \left[ a \ln(a) + b \ln(b) + c \ln(c) + d \ln(d) + \right. \\ \left. N \ln(N) - s \ln(s) - f \ln(f) - m \ln(m) - n \ln(n) \right]$$

### Power

The power of this test is computed using the enumeration procedure described above. When large sample results are needed, the results for the  $z$  test are used.

## T-Test

Based on a study of the behavior of several tests, D'Agostino (1988) and Upton (1982) proposed using the usual two-sample  $t$ -test for testing whether two proportions are equal. One substitutes a '1' for a success and a '0' for a failure in the usual, two-sample  $t$ -test formula. The test statistic is computed as

$$t_{N-2} = (ad - bc) \left( \frac{N-2}{N(nac + mbd)} \right)^{\frac{1}{2}}$$

which can be compared to the  $t$  distribution with  $N-2$  degrees of freedom.

### Power

The power of this test is computed using the enumeration procedure described above, except that the  $t$  tables are used instead of the standard normal tables.



## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are  $P1$ ,  $Beta$ ,  $N1$ , and  $N2$ . 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.

### Test Statistic

Specify which test statistic will be used in searching and reporting.

Note that 'C.C.' is an abbreviation for *Continuity Correction*. This refers to the adding or subtracting  $2/(N1+N2)$  to (or from) the numerator of the z-value to bring the normal approximation closer to the binomial distribution.

### P2 (Control Group Proportion)

Specify the value of  $p_2$ , the control, baseline, or standard group's proportion. The null hypothesis is that the two proportions,  $p_1$  and  $p_2$ , are both equal to this value.

Since these values are proportions, values must be between zero and one.

You may enter a range of values such as  $0.1, 0.2, 0.3$  or  $0.1$  to  $0.9$  by  $0.1$ .

### Alpha (Significance Level)

This option specifies one or more values for the probability of a type-I error (alpha), often called the significance level. This is the probability of a false positive. A type-I error occurs when you reject the null hypothesis of equal proportions when in fact they 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. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

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

### Beta (1 - Power)

This option specifies one or more values for the probability of a type-II error (beta or false negative). A type-II error occurs when you fail to reject the null hypothesis of equal proportions when in fact they are different.

Values must be between zero and one. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.

### N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. 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  $N2 = N1$ .

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions, P1 and P2. P1 is the value of the P1 at which the power is calculated.

### P1 (Group 1 Proportion |H1)

This is the value of P1 under the alternative hypothesis, H1. The power calculations assume that this is the actual value of this proportion.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

Note that values must be between zero and one and cannot be equal to P2.

### Alternative Hypothesis (H1)

This option specifies the type of alternative hypothesis. The null hypothesis is  $H_0: P_1 = P_2$ .

*One-Sided ( $H_1: P_1 < P_2$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H_1: P_1 < P_2$ .

*One-Sided ( $H_1: P_1 > P_2$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H_1: P_1 > P_2$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H_1: P_1 \neq P_2$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1$  is the value of the group 1 proportion at which the power is calculated. Once  $P2$  and  $D1$  are given, the value of  $P1$  can be calculated.

### D1 (Difference|H1 = P1-P2)

This option specifies the difference between the two proportions under the alternative hypothesis,  $H1$ . This difference is used with  $P2$  to calculate the value of  $P1$  using the formula:  $P1 = D1 + P2$ . Differences must be between -1 and 1. They cannot take on the values -1, 0, or 1.

The power calculations assume that  $P1$  is the actual value of the proportion in group 1 (the experimental or treatment group).

You may enter a range of values such as *0.03 0.05 0.10 or 0.01 to 0.05 by 0.01*.

### Alternative Hypothesis (H1)

This option specifies the type of alternative hypothesis. The null hypothesis is  $H0: P1 = P2$ .

*One-Sided ( $H1: D1 < 0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: D1 < 0$ .

*One-Sided ( $H1: D1 > 0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: D1 > 0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: D1 \neq 0$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1$  is the value of the group 1 proportion at which the power is calculated. Once  $P2$  and  $R1$  are given, the value of  $P1$  can be calculated.

### R1 (Ratio|H1 = $P2/P1$ )

This option specifies the ratio between the two proportions,  $P1$  and  $P2$ . This ratio is used with  $P2$  to calculate the value of  $P1$  at which the power is calculated using the formula:  $P1=(R1) \times (P2)$ . The power calculations assume that  $P1$  is the actual value of the proportion in group 1, which is the experimental, or treatment, group.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*. Ratios must be greater than zero. They cannot take on the value of one.

### Alternative Hypothesis (H1)

This option specifies the type of alternative hypothesis. The null hypothesis is  $H0: P1 = P2$ .

*One-Sided ( $H1: R1 < 1$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: R1 < 1$ .

*One-Sided ( $H1: R1 > 1$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: R1 > 1$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: R1 \neq 1$ , Here ' $\neq$ ' means unequal.

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

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratios. Note that the odds are defined as  $O2 = P2 / (1 - P2)$  and  $O1 = P1 / (1 - P1)$ . Once  $P2$  and  $OR1$  are given, the value of  $P1$  can be calculated.

### OR1 (Odds Ratio|H1 = O1/O2)

This option specifies the odds ratio of the two proportions,  $P1$  and  $P2$ . This odds ratio is used with  $P2$  to calculate the value of  $P1$ . The power calculations assume that  $P1$  is the actual value of the proportion in group 1, which is the experimental, or treatment, group.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*. Odds ratios must be greater than zero. They cannot take on the value of one.

### Alternative Hypothesis (H1)

This option specifies the type of alternative hypothesis. The null hypothesis is  $H0: P1 = P2$ .

*One-Sided (H1:OR1<1)* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: P1 < P2$ ,  $H1: D1 < 0$ ,  $H1: R1 < 1$ , or  $H1: OR1 < 1$ .

*One-Sided (H1: OR1>1)* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: P1 > P2$ ,  $H1: D1 > 0$ ,  $H1: R1 > 1$ , or  $H1: OR1 > 1$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: OR1 \neq 1$ . Here ' $\neq$ ' means unequal.

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

## Options Tab

The Options tab contains various limits and options.

### Zero Count Adjustment Method

Zero cell counts often cause calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option *Add to zero cells only*.

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts. We have found that adding 0.0001 seems to work well.

### Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This value indicates the amount that is added. We have found that 0.0001 works well.

Be warned that the value of the ratio and the odds ratio will be affected by the amount specified here!

### Maximum N1 or N2 Exact

When either N1 or N2 is above this amount, power calculations are based on the normal approximation to the binomial. In this case, the actual value of alpha is not calculated. Currently, for three-gigahertz computers, a value near 200 is reasonable. As computers increase in speed, this number may be increased.

### Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

### Calculate Exact Test Results

When checked, the power of Fisher's Exact Test will be calculated for the comparative reports, even if the 'Test Statistic' option is not set to *Fisher's Exact Test*.

This option is provided because calculations for Fisher's Exact Test can become lengthy for large sample sizes.

## Example1 - Finding Power

A study is being designed to study the effectiveness of a new treatment. Historically, the standard treatment has enjoyed a 60% cure rate. Researchers want to compute the power of the two-sided z-test at group sample sizes ranging from 50 to 650 for detecting differences of 0.05 and 0.10 in the cure rate at the 0.05 significance level.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Z Test (Pooled)</b>
D1 .....	<b>0.05 0.10</b>
P2.....	<b>0.6</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>50 to 650 by 100</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>400</b>



## Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 – P2**  
**H0: P1-P2=0. H1: P1-P2<>D1. Test Statistic: Z test with pooled variance**

	Sample Size Grp 1	Sample Size Grp 2	Prop H1 Grp 1 or Trtmnt P1	Prop Grp 2 or Control P2	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
Power	N1	N2	P1	P2					
0.08173	50	50	0.65000	0.60000	0.00000	0.05000	0.05000	0.05239	0.91827
0.14469	150	150	0.65000	0.60000	0.00000	0.05000	0.05000	0.05173	0.85531
0.20852	250	250	0.65000	0.60000	0.00000	0.05000	0.05000	0.04981	0.79148
0.27586	350	350	0.65000	0.60000	0.00000	0.05000	0.05000	0.04946	0.72414
0.34064	450	450	0.65000	0.60000	0.00000	0.05000	0.05000		0.65936
0.40234	550	550	0.65000	0.60000	0.00000	0.05000	0.05000		0.59766
0.46095	650	650	0.65000	0.60000	0.00000	0.05000	0.05000		0.53905
0.18042	50	50	0.70000	0.60000	0.00000	0.10000	0.05000	0.05239	0.81958
0.43968	150	150	0.70000	0.60000	0.00000	0.10000	0.05000	0.05173	0.56032
0.65180	250	250	0.70000	0.60000	0.00000	0.10000	0.05000	0.04981	0.34820
0.79585	350	350	0.70000	0.60000	0.00000	0.10000	0.05000	0.04946	0.20415
0.88326	450	450	0.70000	0.60000	0.00000	0.10000	0.05000		0.11674
0.93640	550	550	0.70000	0.60000	0.00000	0.10000	0.05000		0.06360
0.96636	650	650	0.70000	0.60000	0.00000	0.10000	0.05000		0.03364

Note: exact results based on the binomial were only made when both N1 and N2 were less than 400.

#### Report Definitions

'Power' is the probability of rejecting a false null hypothesis. It should be close to one.

'N1 and N2' are the sizes of the samples drawn from the corresponding populations.

'P1' is the proportion for group one under H1. This is the treatment or experimental group.

'P2' is the proportion for group two. This is the standard, reference, or control group

'D1: Diff. if H1' is the difference P1 – P2 assuming the alternative hypothesis.

'Target Alpha' is the probability of rejecting a true null hypothesis that was desired.

'Actual Alpha' is the value of alpha that is actually achieved.

'Beta' is the probability of accepting a false null hypothesis.

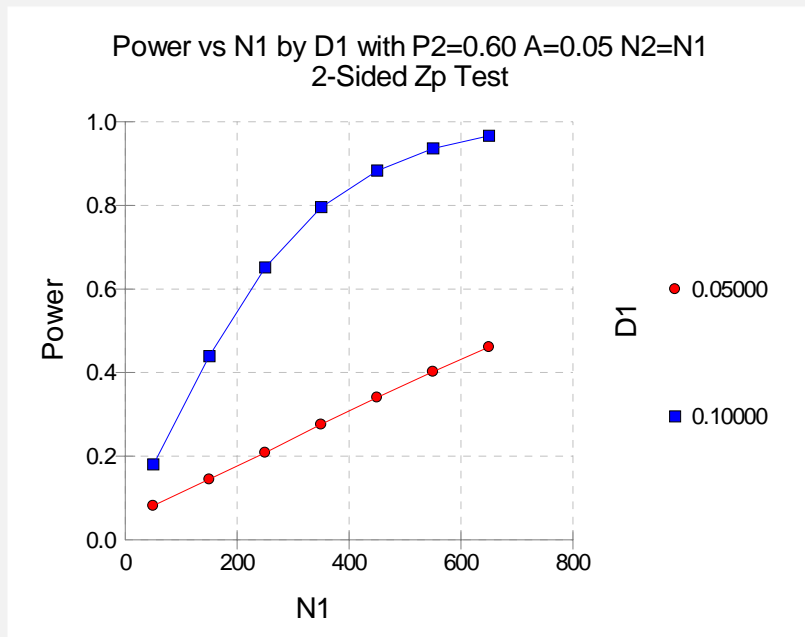
#### Summary Statements

Group sample sizes of 50 in group one and 50 in group two achieve 8% power to detect a difference between the group proportions of 0.05000. The proportion in group one (the treatment group) is assumed to be 0.60000 under the null hypothesis and 0.65000 under the alternative hypothesis. The proportion in group two (the control group) is 0.60000. The test statistic used is the two-sided Z test. The significance level of the test was targeted at 0.05000. The significance level actually achieved by this design is 0.05239.

This report shows the values of each of the parameters, one scenario per row. Note that the actual alpha value is blank for sample sizes greater than 400, which was the limit set for exact computation.

The values from this table are plotted in the chart below.

## Plots Section



## Example2 - Finding the Sample Size

A clinical trial is being designed to test effectiveness of new drug in reducing mortality. Suppose the current cure rate during the first year is 0.44. The sample size should be large enough to detect a difference in the cure rate of 0.10. Assuming the test statistic is a two-sided  $z$ -test with a significance level of 0.05, what sample size will be necessary to achieve 90% power?

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Z Test (Pooled)</b>
D1.....	<b>0.10</b>
P2.....	<b>0.44</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1.....	<b><i>Ignored since this is the search parameter.</i></b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.10</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 – P2**  
**H0: P1-P2=0. H1: P1-P2<>D1. Test Statistic: Z test with pooled variance**

	Sample Size Grp 1	Sample Size Grp 2	Prop H1 Grp 1 or Trtmnt P1	Prop Grp 2 or Control P2	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>							
0.90050	524	524	0.54000	0.44000	0.00000	0.10000	0.05000		0.09950

The required sample size is 524 per group.

These results use the large sample approximation. As an exercise, reset the Maximum N1 or N2 Exact parameter to 600 so that exact results can be calculated. When this is done, the sample size is 521—not much of a difference from the 524 that was found by approximate methods. The actual alpha is 0.04930 which is very close to the target of 0.05.

## Example3 – Comparing the Power of Several Test Statistics

Researchers want to determine which of the eight test statistics to adopt using the comparative reports and charts that *PASS* produces. They want to detect a difference of 0.20 when the response rate of the control group is 0.30. The significance level is 0.05. They want to study sample sizes from 10 to 100.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Z Test (Pooled)</b>
D1 .....	<b>0.2</b>
P2.....	<b>0.3</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>10 to 100 by 10</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>400</b>
Calculate Exact Test Results .....	<b>Checked</b>
<b>Reports Tab</b>	
Show Numeric Report.....	<b>Not checked</b>
Show Comparative Reports .....	<b>Checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots.....	<b>Checked</b>
Summary Statement Rows .....	<b>0</b>

### Annotated Output

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

## Numeric Results

### Power Comparison of Tests Based on the Difference: $P_1 - P_2$

$H_0: P_1 - P_2 = 0$ .  $H_1: P_1 - P_2 \neq 0$ .

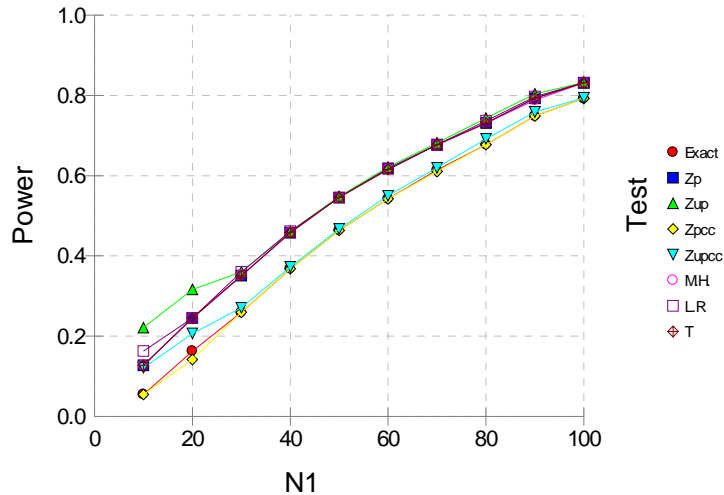
N1/N2	P1	P2	Target Alpha	Exact Test Power	Z(P) Test Power	Z(UnP) Test Power	Z(P) cc Test Power	Z(UnP) cc Test Power	Mantel Hnzl. Power	Like. Ratio Power	T Test Power
10/10	0.5000	0.3000	0.0500	0.0547	0.1275	0.2215	0.0547	0.1215	0.1275	0.1629	0.1275
20/20	0.5000	0.3000	0.0500	0.1632	0.2452	0.3167	0.1419	0.2067	0.2452	0.2452	0.2452
30/30	0.5000	0.3000	0.0500	0.2594	0.3511	0.3604	0.2594	0.2708	0.3511	0.3604	0.3511
40/40	0.5000	0.3000	0.0500	0.3683	0.4581	0.4612	0.3683	0.3728	0.4581	0.4612	0.4581
50/50	0.5000	0.3000	0.0500	0.4635	0.5455	0.5481	0.4635	0.4671	0.5455	0.5455	0.5455
60/60	0.5000	0.3000	0.0500	0.5424	0.6177	0.6214	0.5424	0.5501	0.6157	0.6177	0.6157
70/70	0.5000	0.3000	0.0500	0.6138	0.6771	0.6815	0.6101	0.6195	0.6771	0.6771	0.6771
80/80	0.5000	0.3000	0.0500	0.6773	0.7310	0.7435	0.6773	0.6917	0.7310	0.7368	0.7310
90/90	0.5000	0.3000	0.0500	0.7485	0.7930	0.8036	0.7485	0.7589	0.7882	0.7969	0.7930
100/100	0.5000	0.3000	0.0500	0.7924	0.8320	0.8328	0.7924	0.7942	0.8316	0.8320	0.8316

### Actual Alpha Comparison of Tests Based on the Difference: $P_1 - P_2$

$H_0: P_1 - P_2 = 0$ .  $H_1: P_1 - P_2 = D_1 \neq 0$ .

N1/N2	P1	P2	Target Alpha	Exact Test Alpha	Z(P) Test Alpha	Z(UnP) Test Alpha	Z(P) cc Test Alpha	Z(UnP) cc Test Alpha	Mantel Hnzl. Alpha	Like. Ratio Alpha	T Test Alpha
10/10	0.5000	0.3000	0.0500	0.0119	0.0371	0.0949	0.0119	0.0258	0.0371	0.0771	0.0371
20/20	0.5000	0.3000	0.0500	0.0248	0.0533	0.0686	0.0214	0.0267	0.0533	0.0534	0.0533
30/30	0.5000	0.3000	0.0500	0.0261	0.0487	0.0583	0.0261	0.0321	0.0487	0.0583	0.0487
40/40	0.5000	0.3000	0.0500	0.0282	0.0484	0.0541	0.0276	0.0317	0.0484	0.0541	0.0484
50/50	0.5000	0.3000	0.0500	0.0307	0.0498	0.0554	0.0307	0.0334	0.0498	0.0498	0.0498
60/60	0.5000	0.3000	0.0500	0.0308	0.0525	0.0552	0.0308	0.0353	0.0483	0.0525	0.0491
70/70	0.5000	0.3000	0.0500	0.0330	0.0516	0.0549	0.0318	0.0348	0.0516	0.0516	0.0516
80/80	0.5000	0.3000	0.0500	0.0331	0.0513	0.0518	0.0331	0.0350	0.0493	0.0516	0.0493
90/90	0.5000	0.3000	0.0500	0.0344	0.0497	0.0525	0.0344	0.0365	0.0497	0.0500	0.0497
100/100	0.5000	0.3000	0.0500	0.0348	0.0510	0.0529	0.0348	0.0373	0.0494	0.0517	0.0494

Power vs N1 by Test with  $D_1=0.20$   $P_2=0.30$   $A=0.05$   
 $N_2=N_1$  2-Sided Test



It is interesting to note that the power of Fisher's Exact Test and the  $z$ -test with continuity correction are consistently lower than the other tests. This occurs because the actual alpha achieved by these tests is much lower than that of the other tests. An interesting finding of this short study was that the regular  $t$ -test performed better than the more popular  $z$ -test.

## Example4 - Validation using Fleiss with Equal Sample Sizes

Fleiss (2003), page 74, presents a sample size study in which  $P1 = 0.7$ ,  $P2 = 0.6$ ,  $\alpha = 0.01$ , and  $\beta = 0.05$ . Assuming two-sided testing and equal sample allocation, Fleiss finds the necessary sample size to be 827 in each group. The calculations of Fleiss (2003) included an adjustment for continuity correction. This continuity correction is not necessary here when exact calculations are made. However, when the sample size is large enough so that approximate calculations are used, the continuity correction must be applied to obtain the same results. This is done by setting the Test Statistic to 'Z Test C.C.'. Note that this adjustment is used here to keep our results identical to those of Fleiss (2003). In practice, this adjustment is not recommended because it reduces the power and the actual alpha of the test procedure.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic .....	<b>Z Test C.C. (Pooled)</b>
D1 .....	<b>0.10</b>
P2.....	<b>0.60</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b><i>Ignored since this is the search parameter.</i></b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.01</b>
Beta.....	<b>0.05</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference:  $P1 - P2$**   
**H0:  $P1-P2=0$ . H1:  $P1-P2 \neq 0$ . Test Statistic: Z test with continuity correction and pooled variance**

	Sample Size	Sample Size	Prop H1 Grp 1 or Trtmnt P1	Prop Grp 2 or Control P2	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>							
0.95025	827	827	0.70000	0.60000	0.00000	0.10000	0.01000		0.04975

**PASS** found the required sample size to be 827 which corresponds to Fleiss.

## Example5 - Validation using Fleiss with Unequal Sample Sizes

Fleiss (2003), pages 76-77, presents a sample size study in which  $P1 = 0.25$ ,  $P2 = 0.40$ ,  $\alpha = 0.01$ , and  $\beta = 0.05$ . Assuming two-sided testing with half as many in the second group as the first, Fleiss finds the sample sizes to be 530 in the first group and 265 in the second.

Note that half as many in the second group is achieved by setting  $R$  to 0.5.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Z Test C.C. (Pooled)</b>
D1.....	<b>-0.15</b>
P2.....	<b>0.40</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1.....	<b><i>Ignored since this is the search parameter.</i></b>
N2.....	<b>Use R</b>
R.....	<b>0.5</b>
Alpha .....	<b>0.01</b>
Beta.....	<b>0.05</b>

### Options Tab

Maximum N1 or N2 Exact ..... **100 (Set low for a rapid search.)**

### Annotated Output

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

### Numeric Results

Numeric Results of Tests Based on the Difference: P2 - P1									
H0: P2-P1=0. H1: P2-P1<>D1. Test Statistic: Z test with continuity correction and pooled variance									
	Sample Size Grp 1	Sample Size Grp 2	Prop H1 Trtmnt P1	Prop Grp 2 or Control P2	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
Power	N1	N2							
0.95066	531	266	0.25000	0.40000	0.00000	-0.15000	0.01000		0.04934

**PASS** found the required sample sizes to be 531 and 266 which nearly corresponds to Fleiss's results. Fleiss computed 530 instead of 531. 531 is correct because the power for 530 is slightly less than the required 0.95.

## Example6 - Determining the Power after Completing an Experiment

A study has just been completed aimed at determining the effectiveness of a new treatment for cancer. Because of the cost of administering the new treatment, they would adopt the new treatment only if the difference between the proportion cured by the new treatment and that cured by the standard treatment is at least 0.10. The researchers enrolled 200 cancer patients in the study (100 for each treatment) and found that 51% were cured by the standard treatment, while 62% were cured by the new treatment. These results, however, showed no statistically significant difference based on the pooled z-test with continuity correction and  $\alpha = 0.05$ . Therefore, the researchers want to compute the power of this test for detecting a difference of 0.10 for standard treatment proportions ranging from 0.40 to 0.60.

Note that the power was not exclusively computed at the observed sample proportion for the standard treatment group, 0.51. It is more informative to compute the power for a range of likely values suggested by historical evidence.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Z Test C.C. (Pooled)</b>
D1 .....	<b>0.10</b>
P2 .....	<b>0.40 to 0.60 by 0.04</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>100</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>400</b>



## Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 - P2**

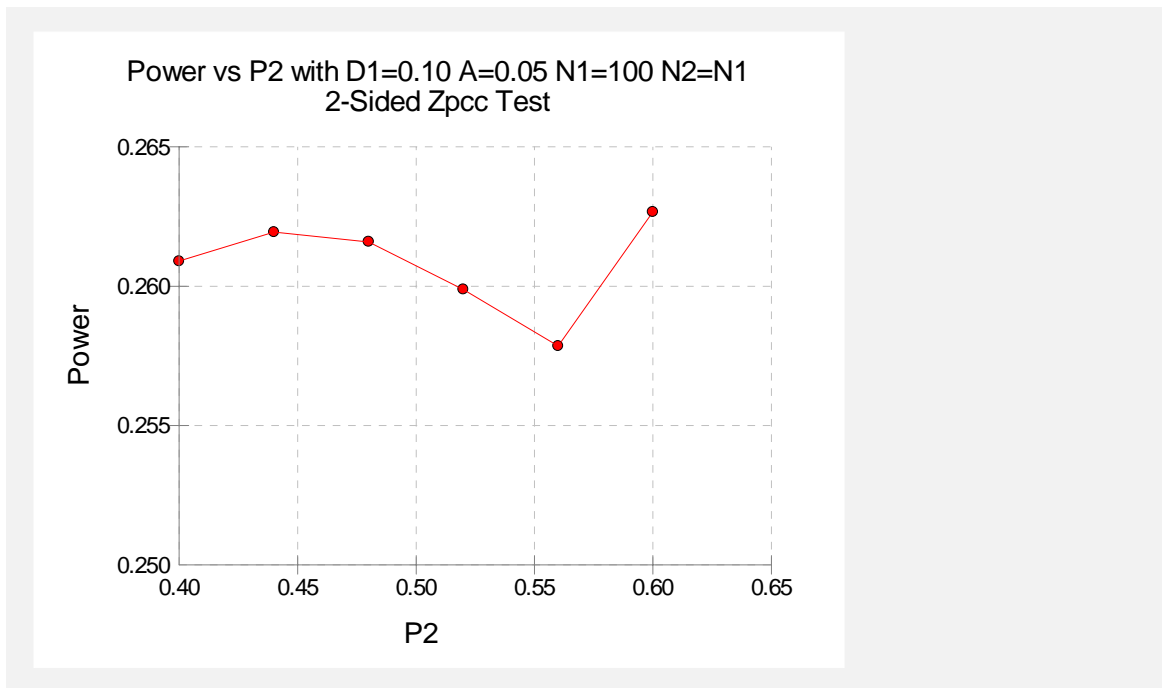
H0: P1-P2=0. H1: P1-P2=D1<>0. Test Statistic: Z test with continuity correction and pooled variance

Power	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop H1 Grp 1 or Trtmnt P1	Prop Grp 2 or Control P2	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
0.26090	100	100	0.50000	0.40000	0.00000	0.10000	0.05000	0.03628	0.73910
0.26194	100	100	0.54000	0.44000	0.00000	0.10000	0.05000	0.03861	0.73806
0.26159	100	100	0.58000	0.48000	0.00000	0.10000	0.05000	0.03988	0.73841
0.25988	100	100	0.62000	0.52000	0.00000	0.10000	0.05000	0.03988	0.74012
0.25785	100	100	0.66000	0.56000	0.00000	0.10000	0.05000	0.03861	0.74215
0.26266	100	100	0.70000	0.60000	0.00000	0.10000	0.05000	0.03628	0.73734

This report shows the values of each of the parameters, one scenario per row. The power over the entire range of the likely standard treatment proportions is relatively constant at 0.26.

The values from this table are plotted in the chart below.

### Plots Section



It is evident from these results that the test performed by the researchers had very low power to detect a difference of 0.10 with the sample size used. The power is only 0.26 for a large range of standard treatment proportions. Note that the fluctuation in power is related to the value of alpha.

## Example7 - Finding the Sample Size using Ratios

Researchers would like to design an experiment to compare the infection rate of a rare disease among two populations. More specifically, they would like to determine how many subjects they need to sample from each population to determine if the disease rate in population 1 is at least three times that of population 2 with 80% power. Suppose that the researchers are confident from previous studies that the infection rate in population 2 is 0.025. The researchers plan to use the likelihood ratio test and  $\alpha = 0.05$ .

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic .....	<b>Likelihood Ratio Test</b>
R1 .....	<b>3</b>
P2.....	<b>0.025</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<i>Ignored since this is the Find setting</i>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Tests Based on the Ratio: P1 / P2  
H0: P1/P2=1. H1: P1/P2=R1<>1. Test Statistic: Likelihood Ratio test

	Sample Size Grp 1	Sample Size Grp 2	Prop H1 Grp 1 or Trtmnt P1	Prop Grp 2 or Control P2	Ratio if H0 R0	Ratio if H1 R1	Target Alpha	Actual Alpha	Beta
Power	N1	N2							
0.8012	298	298	0.0750	0.0250	1.000	3.000	0.0500		0.1988

The researchers must sample 298 individuals from each population to achieve 80% power to detect a ratio of 3.0.

## Chapter 205

# Two Independent Proportions – Offset

## Introduction

This module computes power and sample size for hypothesis tests of the difference, ratio, or odds ratio of two independent proportions. The word 'offset' in the chapter title indicates that the difference being tested in the null hypothesis is not zero (or that the ratio is not one). The *non-offset* case is available in another procedure. This procedure compares the power achieved by each of several test statistics.

The power calculations assume that independent, random samples are drawn from two populations.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Technical Details

Suppose you have two populations from which dichotomous (binary) responses will be recorded. The probability (or risk) of obtaining the event of interest in population 1 (the treatment group) is  $p_1$  and in population 2 (the control group) is  $p_2$ . The corresponding failure proportions are given by  $q_1 = 1 - p_1$  and  $q_2 = 1 - p_2$ .

An assumption is made that the responses from each group follow a binomial distribution. This means that the event probability,  $p_i$ , is the same for all subjects within the group and that the response from one subject is independent of that of any other subject.

Random samples of  $m$  and  $n$  individuals are obtained from these two populations. The data from these samples can be displayed in a 2-by-2 contingency table as follows

Group	Success	Failure	Total
Treatment	$a$	$c$	$m$
Control	$b$	$d$	$n$
Total	$s$	$f$	$N$

The following alternative notation is sometimes used.

Group	Success	Failure	Total
Treatment	$x_{11}$	$x_{12}$	$n_1$
Control	$x_{21}$	$x_{22}$	$n_2$
Total	$m_1$	$m_2$	$N$

The binomial proportions,  $p_1$  and  $p_2$ , are estimated from these data using the formulae

$$\hat{p}_1 = \frac{a}{m} = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{b}{n} = \frac{x_{21}}{n_2}$$

## Comparing Two Proportions

When analyzing studies such as this, you usually want to compare the two binomial probabilities,  $p_1$  and  $p_2$ . The most direct method of comparing these quantities is to calculate their difference or their ratio. If the binomial probability is expressed in terms of odds rather than probability, another measure is the odds ratio. Mathematically, these comparison parameters are

<u>Parameter</u>	<u>Computation</u>
Difference	$\delta = p_1 - p_2$
Risk Ratio	$\phi = p_1 / p_2$
Odds Ratio	$\psi = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 q_2}{p_2 q_1}$

The choice of which of these measures is used might seem arbitrary, but it is not. Not only will the interpretation be different, but, for small sample sizes, the powers of tests based on different parameters will be different. The non-null case is commonly used in equivalence and non-inferiority testing.

### Difference

The (risk) difference,  $\delta = p_1 - p_2$ , is perhaps the most direct method of comparison between the two event probabilities. This parameter is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One interpretation difficulty occurs when the event of interest is rare. If a difference of 0.001 were reported for an event with a baseline probability of 0.40, we would probably dismiss this as being of little importance. That is, there usually little interest in a treatment that decreases the probability from 0.400 to 0.399. However, if the baseline probability of a disease was 0.002 and 0.001 was the decrease in the disease probability, this would represent a reduction of 50%. Thus we see that interpretation depends on the baseline probability of the event.

A similar situation occurs when the amount of possible difference is considered. Consider two events, one with a baseline event rate of 0.40 and the other with a rate of 0.02. What is the maximum decrease that can occur? Obviously, the first event rate can be decreased by an absolute amount of 0.40, while the second can only be decreased by a maximum of 0.02.

So, although creating the simple difference is a useful method of comparison, care must be taken that it is appropriate for the situation.

## Ratio

The (risk) ratio,  $\phi = p_1 / p_2$ , gives the relative change in the disease risk due to the application of the treatment. This parameter is also direct and easy to interpret. To compare this with the difference, consider a treatment that reduces the risk of disease from 0.1437 to 0.0793. One should consider which single number is more enlightening, the fact that the absolute risk of disease has been decreased by 0.0644, or the fact that risk of disease in the treatment group is only 55.18% of that in the control group. In many cases, the percentage (risk ratio) communicates the impact of the treatment better than the absolute change.

Perhaps the biggest drawback to this parameter is that it cannot be calculated in one of the most common experimental designs, the case-control study. Another drawback is that the odds ratio occurs directly in the likelihood equations and as a parameter in logistic regression.

## Odds Ratio

Chances are usually communicated as long-term proportions or probabilities. In betting, chances are often given as odds. For example, the odds of a horse winning a race might be set at 10-to-1 or 3-to-2. Odds can easily be translated into probability. An odds of 3-to-2 means that the event is expected to occur three out of five times. That is, an odds of 3-to-2 (1.5) translates to a probability of winning of 0.60.

The odds of an event are calculated by dividing the event risk by the non-event risk. Thus, in our case of two populations, the odds are

$$o_1 = \frac{p_1}{1 - p_1} \text{ and } o_2 = \frac{p_2}{1 - p_2}$$

For example, if  $p_1$  is 0.60, the odds are  $0.60/0.40 = 1.5$ . Rather than represent the odds as a decimal amount, it is re-scaled into whole numbers. Instead of saying the odds are 1.5-to-1, we say they are 3-to-2.

Another way to compare proportions is to compute the ratio of their odds. The odds ratio of two events is

$$\psi = \frac{o_1}{o_2} = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 q_2}{p_2 q_1}$$

Although the odds ratio is more complicated to interpret than the risk ratio, it is often the parameter of choice. One reason for this is the fact that the odds ratio can be accurately estimated from case-control studies, while the risk ratio cannot. Also, the odds ratio is the basis of logistic regression (used to study the influence of risk factors). Furthermore, the odds ratio is the natural parameter in the conditional likelihood of the two-group, binomial-response design. Finally, when the baseline event rates are rare, the odds ratio provides a close approximation to the risk ratio since, in this case,  $1 - p_1 \approx 1 - p_2$ , so that

$$\psi = \frac{o_1}{o_2} = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \approx \frac{p_1}{p_2} = \phi$$

# Hypothesis Tests

Although several statistical tests are available for testing the inequality of two proportions, only a few can be generalized to the non-null case. No single test is the champion in every situation, so one should compare the powers of the various tests to determine which to use.

## Difference

The (risk) difference,  $\delta = p_1 - p_2$ , is perhaps the most direct method for comparing two proportions. Three sets of statistical hypotheses can be formulated:

1.  $H_0: p_1 - p_2 = \delta_0$  versus  $H_1: p_1 - p_2 \neq \delta_0$ ; this is often called the *two-tailed test*.
2.  $H_0: p_1 - p_2 \leq \delta_0$  versus  $H_1: p_1 - p_2 > \delta_0$ ; this is often called the *upper-tailed test*.
3.  $H_0: p_1 - p_2 \geq \delta_0$  versus  $H_1: p_1 - p_2 < \delta_0$ ; this is often called the *lower-tailed test*.

## Ratio

The (risk) ratio,  $\phi = p_1 / p_2$ , is often preferred as a comparison parameter because it expresses the difference as a percentage rather than an amount. Three sets of statistical hypotheses can be formulated:

1.  $H_0: p_1 / p_2 = \phi_0$  versus  $H_1: p_1 / p_2 \neq \phi_0$ ; this is often called the *two-tailed test*.
2.  $H_0: p_1 / p_2 \leq \phi_0$  versus  $H_1: p_1 / p_2 > \phi_0$ ; this is often called the *upper-tailed test*.
3.  $H_0: p_1 / p_2 \geq \phi_0$  versus  $H_1: p_1 / p_2 < \phi_0$ ; this is often called the *lower-tailed test*.

## Odds Ratio

The odds ratio,  $\psi = [p_1 / (1 - p_1)] / [p_2 / (1 - p_2)]$ , is sometimes used as the comparison because of its statistical properties and because some convenient experimental designs only allow it to be estimated. Three sets of statistical hypotheses can be formulated:

1.  $H_0: \psi = \psi_0$  versus  $H_1: \psi \neq \psi_0$ ; this is often called the *two-tailed test*.
2.  $H_0: \psi \leq \psi_0$  versus  $H_1: \psi > \psi_0$ ; this is often called the *upper-tailed test*.
3.  $H_0: \psi \geq \psi_0$  versus  $H_1: \psi < \psi_0$ ; this is often called the *lower-tailed test*.

## Power Calculation

The power for a test statistic that is based on the normal approximation can be computed exactly using two binomial distributions. The following steps are taken to compute the power of such a test.

1. Find the critical value (or values in the case of a two-sided test) using the standard normal distribution. The critical value,  $z_{critical}$ , is that value of  $z$  that leaves exactly the target value of alpha in the appropriate tail of the normal distribution. For example, for an upper-tailed test with a target alpha of 0.05, the critical value is 1.645.
2. Compute the value of the test statistic,  $z_t$ , for every combination of  $x_{11}$  and  $x_{21}$ . Note that  $x_{11}$  ranges from 0 to  $n_1$ , and  $x_{21}$  ranges from 0 to  $n_2$ . A small value (around 0.0001) can be added to the zero cell counts to avoid numerical problems that occur when the cell value is zero.
3. If  $z_t > z_{critical}$ , the combination is in the rejection region. Call all combinations of  $x_{11}$  and  $x_{21}$  that lead to a rejection the set  $A$ .
4. Compute the power for given values of  $p_1$  and  $p_2$  as

$$1 - \beta = \sum_A \binom{n_1}{x_{11}} p_1^{x_{11}} q_1^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_2^{x_{21}} q_2^{n_2 - x_{21}}$$

5. Compute the actual value of alpha achieved by the design by substituting  $p_2$  for  $p_1$  to obtain

$$\alpha^* = \sum_A \binom{n_1}{x_{11}} \binom{n_2}{x_{21}} p_2^{x_{11} + x_{21}} q_2^{n_1 + n_2 - x_{11} - x_{21}}$$

## Asymptotic Approximations

When the values of  $n_1$  and  $n_2$  are large (say over 200), these formulas often take a long time to evaluate. In this case, a large sample approximation is used. The large sample approximation is made by replacing the values of  $\hat{p}_1$  and  $\hat{p}_2$  in the  $z$  values with the corresponding values of  $p_1$  and  $p_2$  under the alternative hypothesis and then computing the results based on the normal distribution. Note that in large samples, the Farrington and Manning statistic is substituted for the Gart and Nam statistic. Also, for large samples, the results for the odds ratio have not (to our knowledge) been published. In this case, we substitute the calculations based on the ratio.



## Test Statistics

Several test statistics have been proposed for testing whether the difference, ratio, or odds ratio are different from a specified value. The main difference among the several test statistics is in the formula used to compute the standard error used in the denominator. These tests are based on the following  $z$ -test

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - c}{\hat{\sigma}}$$

The constant,  $c$ , represents a continuity correction that is applied in some cases. When the continuity correction is not used,  $c$  is zero. In power calculations, the values of  $\hat{p}_1$  and  $\hat{p}_2$  are not known. The corresponding values of  $p_1$  and  $p_2$  under the alternative hypothesis are reasonable substitutes.

Following is a list of the test statistics available in **PASS**. The availability of several test statistics begs the question of which test statistic you should use. The answer is simple: you should use the test statistic that you will use to analyze your data. You may choose a method because it is a standard in your industry, because it seems to have better statistical properties, or because your statistical package calculates it. Whatever your reasons for selecting a certain test statistic, you should use the same test statistic during power or sample calculations.

### Z Test (Pooled)

This test was first proposed by Karl Pearson in 1900. Although this test is usually expressed directly as a chi-square statistic, it is expressed here as a  $z$  statistic so that it can be more easily used for one-sided hypothesis testing. The proportions are pooled (averaged) in computing the standard error. The formula for the test statistic is

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_1}$$

where

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

### Z Test (Unpooled)

This test statistic does not pool the two proportions in computing the standard error.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_2}$$

where

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

### Z Test with Continuity Correction (Pooled)

This test is the same as Z Test (Pooled), except that a continuity correction is used. Recall that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 + \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_1}$$

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

where  $F$  is -1 for lower-tailed, 1 for upper-tailed, and both -1 and 1 for two-sided hypotheses.

### Z Test with Continuity Correction (Unpooled)

This test is the same as the Z Test (Unpooled), except that a continuity correction is used. Recall that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_2}$$

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

where  $F$  is -1 for lower-tailed, 1 for upper-tailed, and both -1 and 1 for two-sided hypotheses.

### T-Test of Difference

Based on a detailed, comparative study of the behavior of several tests, D'Agostino (1988) and Upton (1982) proposed using the usual two-sample  $t$ -test for testing whether the two proportions are equal. One substitutes a '1' for a success and a '0' for a failure in the usual, two-sample  $t$ -test formula.

## Miettinen and Nurminen's Likelihood Score Test of the Difference

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the difference is equal to a specified, non-zero, value,  $\delta_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing this test statistic is

$$z_{MND} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_{MND}}$$

where

$$\hat{\sigma}_{MND} = \sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right) \left( \frac{N}{N-1} \right)}$$

$$\tilde{p}_1 = \tilde{p}_2 + \delta_0$$

$$\tilde{p}_2 = 2B \cos(A) - \frac{L_2}{3L_3}$$

$$A = \frac{1}{3} \left[ \pi + \cos^{-1} \left( \frac{C}{B^3} \right) \right]$$

$$B = \text{sign}(C) \sqrt{\frac{L_2^2}{9L_3} - \frac{L_1}{3L_3}}$$

$$C = \frac{L_2^3}{27L_3^3} - \frac{L_1 L_2}{6L_3^2} + \frac{L_0}{2L_3}$$

$$L_0 = x_{21} \delta_0 (1 - \delta_0)$$

$$L_1 = [N_2 \delta_0 - N - 2x_{21}] \delta_0 + M_1$$

$$L_2 = (N + N_2) \delta_0 - N - M_1$$

$$L_3 = N$$

### Miettinen and Nurminen's Likelihood Score Test of the Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the ratio is equal to a specified value,  $\phi_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 / \tilde{p}_2 = \phi_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{MNR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \phi_0^2 \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right) \left( \frac{N}{N-1} \right)}}$$

where

$$\begin{aligned}\tilde{p}_1 &= \tilde{p}_2 \phi_0 \\ \tilde{p}_2 &= \frac{-B - \sqrt{B^2 - 4AC}}{2A} \\ A &= N\phi_0 \\ B &= -[N_1\phi_0 + x_{11} + N_2 + x_{21}\phi_0] \\ C &= M_1\end{aligned}$$

### Miettinen and Nurminen's Likelihood Score Test of the Odds Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the odds ratio is equal to a specified value,  $\psi_0$ . Because the approach they used with the difference and ratio does not easily extend to the odds ratio, they used a score statistic approach for the odds ratio. The regular MLE's are  $\hat{p}_1$  and  $\hat{p}_2$ . The constrained MLE's are  $\tilde{p}_1$  and  $\tilde{p}_2$ . These estimates are constrained so that  $\tilde{\psi} = \psi_0$ . A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{MNO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{N_2 \tilde{p}_1 \tilde{q}_1} + \frac{1}{N_2 \tilde{p}_2 \tilde{q}_2} \right) \left( \frac{N}{N-1} \right)}}$$

where

$$\begin{aligned}\tilde{p}_1 &= \frac{\tilde{p}_2 \psi_0}{1 + \tilde{p}_2 (\psi_0 - 1)}, \quad \tilde{p}_2 = \frac{-B + \sqrt{B^2 - 4AC}}{2A}, \\ A &= N_2(\psi_0 - 1), \quad B = N_1\psi_0 + N_2 - M_1(\psi_0 - 1), \quad C = -M_1\end{aligned}$$

## Farrington and Manning's Likelihood Score Test of the Difference

Farrington and Manning (1990) proposed a test statistic for testing whether the difference is equal to a specified value,  $\delta_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ , are used in the denominator. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMD} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right)}}$$

where the estimates,  $\tilde{p}_1$  and  $\tilde{p}_2$ , are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

## Farrington and Manning's Likelihood Score Test of the Ratio

Farrington and Manning (1990) proposed a test statistic for testing whether the ratio is equal to a specified value,  $\phi_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 / \tilde{p}_2 = \phi_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to increase the variance estimate. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \phi_0^2 \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right)}}$$

where the estimates,  $\tilde{p}_1$  and  $\tilde{p}_2$ , are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

## Farrington and Manning's Likelihood Score Test of the Odds Ratio

Farrington and Manning (1990) indicate that the Miettinen and Nurminen statistic may be modified by removing the factor  $N/(N-1)$ .

The formula for computing this test statistic is

$$z_{FMO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{N_2 \tilde{p}_1 \tilde{q}_1} + \frac{1}{N_2 \tilde{p}_2 \tilde{q}_2} \right)}}$$

where the estimates,  $\tilde{p}_1$  and  $\tilde{p}_2$ , are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

### Gart and Nam's Likelihood Score Test of the Difference

Gart and Nam (1990), page 638, proposed a modification to the Farrington and Manning (1988) difference test that corrected for skewness. Let  $z_{FMD}(\delta)$  stand for the Farrington and Manning difference test statistic described above. The skewness corrected test statistic,  $z_{GND}$ , is the appropriate solution to the quadratic equation

$$(-\tilde{\gamma})z_{GND}^2 + (-1)z_{GND} + (z_{FMD}(\delta) + \tilde{\gamma}) = 0$$

where

$$\tilde{\gamma} = \frac{\tilde{V}^{3/2}(\delta)}{6} \left( \frac{\tilde{p}_1 \tilde{q}_1 (\tilde{q}_1 - \tilde{p}_1)}{n_1^2} - \frac{\tilde{p}_2 \tilde{q}_2 (\tilde{q}_2 - \tilde{p}_2)}{n_2^2} \right)$$

### Gart and Nam's Likelihood Score Test of the Ratio

Gart and Nam (1988), page 329, proposed a modification to the Farrington and Manning (1988) ratio test that corrected for skewness. Let  $z_{FMR}(\phi)$  stand for the Farrington and Manning ratio test statistic described above. The skewness corrected test statistic,  $z_{GNR}$ , is the appropriate solution to the quadratic equation

$$(-\tilde{\varphi})z_{GNR}^2 + (-1)z_{GNR} + (z_{FMR}(\phi) + \tilde{\varphi}) = 0$$

where

$$\tilde{\varphi} = \frac{1}{6\tilde{u}^{3/2}} \left( \frac{\tilde{q}_1 (\tilde{q}_1 - \tilde{p}_1)}{n_1^2 \tilde{p}_1^2} - \frac{\tilde{q}_2 (\tilde{q}_2 - \tilde{p}_2)}{n_2^2 \tilde{p}_2^2} \right)$$

$$\tilde{u} = \frac{\tilde{q}_1}{n_1 \tilde{p}_1} + \frac{\tilde{q}_2}{n_2 \tilde{p}_2}$$

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *P1.1*, *Alpha*, *Beta*, *N1*, and *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.

### Test Statistic

Specify which test statistic is used in searching and reporting.

Note that *C.C.* is an abbreviation for *Continuity Correction*. This refers to the adding or subtracting  $1/(2n)$  to (or from) the numerator of the z-value to bring the normal approximation closer to the binomial distribution.

## P2 (Control Group Proportion)

Specify the value of P2, the control, baseline, or standard group's proportion. The null hypothesis is that the two proportions differ by a specified amount. Since P2 is a proportion, these values must be between zero and one.

You may enter a range of values such as *0.1,0.2,0.3* or *0.1 to 0.9 by 0.1*.

## N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. 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  $N2 = N1$ .

## 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 that the proportions differ by a specified amount when in fact they do differ by that amount.

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. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

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

## 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 proportions when in fact they are different.

Values must be between zero and one. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.



## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions,  $P1$  and  $P2$ .  $P1.0$  is the value of the  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated.

### P1.0 (Group 1 Proportion|H0)

This option specifies the value of the group 1 proportion given the null hypothesis. The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. In this non-null case, the value of  $P1.0$  is not equal to  $P2$  as it is in the null case.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

Proportions must be between zero and one. They cannot take on the values zero or one.

### P1.1 (Group 1 Proportion|H1)

This is the value of  $P1$  under the alternative hypothesis. It is written  $P1.1$ . The power calculations assume that this is the actual value of this proportion.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

Note that values must be between zero and one.

## Alternative Hypothesis (H1)

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided ( $H1: P1 < P2$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: P1 < P2$ .

*One-Sided ( $H1: P1 > P2$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: P1 > P2$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: P1 \neq P2 + D0$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $D0$ , and  $D1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### D0 (Difference|H0 = P1.0 – P2)

This option specifies the difference between the two proportions given in the null hypothesis,  $H0$ . This difference is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = P2 + D0$ . Note that  $P1.0$  here means the value of  $P1$  under  $H0$ .

Differences must be between -1 and 1. They cannot take on the values -1, 0, or 1.

The power calculations use  $P1.0$  as the value of the proportion in group 1 (the experimental or treatment group) under the null hypothesis. In this non-null case, the value of  $P1.0$  is not equal to  $P2$  as it is in the null case.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

### D1 (Difference|H0 = P1.1 – P2)

This option specifies the difference between the  $P1.1$  and  $P2$ . This difference is used with  $P2$  to calculate the value of  $P1.1$  using the formula:  $P1.1 = D1 + P2$ . Note that  $P1.1$  here means the value of  $P1$  under  $H1$ . Differences must be between -1 and 1. They cannot take on the values -1 or 1.

The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group).

This option is only used if you are specifying *Differences*.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

## Alternative Hypothesis (H1)

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided (H1:D1<D0)* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: D1 < D0$ .

*One-Sided (H1:D1>D0)* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: D1 > D0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: D1 \neq D0$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $R0$ , and  $R1$  are given, the values of  $P1.0$  and  $P1.1$  can be calculated.

### **R0 (Ratio|H0 = $P1.0 / P2$ )**

This option specifies the ratio between the group 1 proportion under the null hypothesis  $P1.0$  and  $P2$ . This ratio is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = R0 \times P2$ . The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. In this non-null case, the value of  $P1.0$  is not equal to  $P2$  as it is in the null case.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*.

Ratios must greater than zero.

### **R1 (Ratio|H1 = $P1.1 / P2$ )**

This option specifies the ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the proportion in group 1 under the alternative hypothesis. This ratio is used with  $P2$  to calculate the value of  $P1.1$  using the formula:  $P1.1 = R1 \times P2$ . The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group).

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*.

Ratios must greater than zero. They cannot take on the value of one.

## **Alternative Hypothesis (H1)**

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided ( $H1: R1 < R0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: R1 < R0$ .

*One-Sided ( $H1: R1 > R0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: R1 > R0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: R1 \neq R0$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratios,  $O1.1 / O2$  and  $O1.0 / O2$ . Note that the odds are defined as  $O2 = P2 / (1 - P2)$ ,  $O1.0 = P1.0 / (1 - P1.0)$ , etc.  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $OR0$ , and  $OR1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### OR0 (Odds Ratio|H0 = O1.0 / O2)

This option specifies the odds ratio between the group 1 proportion under the null hypothesis and the proportion in group 2. This value is used with  $P2$  to calculate the value of  $P1.0$ . The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. In this non-null case, the value of  $P1.0$  is not equal to  $P2$  as it is in the null case.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*.

Odds ratios must greater than zero.

### OR1 (Odds Ratio|H1 = O1.1 / O2)

This option specifies the odds ratio of the two proportions  $P1.1$  and  $P2$ . This odds ratio is used with  $P2$  to calculate the value of  $P1.1$ . The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1, which is the experimental, or treatment, group.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*. Odds ratios must greater than zero.

## Alternative Hypothesis (H1)

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided (H1:  $OR1 < OR0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: OR1 < OR0$ .

*One-Sided (H1:  $OR1 > OR0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: OR1 > OR0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: OR1 \neq OR0$ . Here ' $\neq$ ' means unequal.

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

# Options Tab

The Options tab contains various limits and options.

## Maximum N1 or N2 Exact

When either N1 or N2 is above this amount, power calculations are based on the normal approximation to the binomial. In this case, the actual value of alpha is not calculated. Currently, for three-gigahertz computers, a value near 200 is reasonable. As computers increase in speed, this number may be increased.

## Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

## Zero Count Adjustment Method

Zero cell counts often cause calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option 'Add to zero cells only'.

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts.. We have found that adding 0.0001 seems to work well.

## Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This value indicates the amount that is added. We have found that 0.0001 works well.

Be warned that the value of the ratio and the odds ratio will be affected by the amount specified here!

## Example1 - Finding Power

A study is being designed to study the effectiveness of a new treatment. Historically, the standard treatment has enjoyed a 60% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the standard treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the standard treatment. The researchers will recommend adoption of the new treatment if it has a cure rate of at least 55%.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data. They want to study the power of the one-sided Farrington and Manning test at group sample sizes ranging from 50 to 2000 for detecting a difference significantly greater than -0.05 when the actual cure rate of the new treatment ranges from 57% to 70%. The significance level will be 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>-0.05</b>
D1 .....	<b>-0.03 0.00 0.05 0.10</b>
P2.....	<b>0.6</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
N1 .....	<b>50 100 250 500 1000 1500 2000</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>300</b>

## Annotated Output

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

### Numeric Results

#### Numeric Results of Tests Based on the Difference: P1 – P2

H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
Power	N1	N2								
0.0778	50	50	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500	0.0527	0.9222
0.0865	100	100	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500	0.0499	0.9135
0.1189	250	250	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500	0.0516	0.8811
0.1583	500	500	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500		0.8417
0.2310	1000	1000	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500		0.7690
0.2976	1500	1500	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500		0.7024
0.3596	2000	2000	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500		0.6404
Report continues ...										

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 300.

#### Report Definitions

'H0' as an abbreviation for the NULL hypothesis. This is the hypothesis being evaluated by the statistical test.

'H1' as an abbreviation for the ALTERNATIVE hypothesis. This hypothesis gives the 'true' parameter values.

'Power' is the probability of rejecting a false null hypothesis. It should be close to one.

'N1 and N2' are the sizes of the samples drawn from the corresponding populations.

'P2' is the proportion for group two. This is the standard, reference, baseline, or control group.

'P1.0' is the proportion for group one (treatment group) assuming the null hypothesis (H0).

'P1.1' is the proportion for group one (treatment group) assuming the alternative hypothesis (H1).

'D0: Diff|H0' is the difference P1 – P2 assuming the null hypothesis (H0).

'D1: Diff|H1' is the difference P1 – P2 assuming the alternative hypothesis (H1).

'Target Alpha' is the probability of rejecting a true null hypothesis that was desired.

'Actual Alpha' is the value of alpha that is actually achieved.

'Beta' is the probability of accepting a false H0. Beta = 1 - Power.

#### Summary Statements

Group sample sizes of 50 in group one and 50 in group two achieve 8% power to detect a difference between the group proportions of -0.0300. The proportion in group two is 0.6000. The proportion in group one is assumed to be 0.5500 under the null hypothesis and 0.5700 under the alternative hypothesis. The test statistic used is the one-sided Score test (Farrington & Manning). The significance level of the test was targeted at 0.0500. The significance level actually achieved by this design is 0.0527.

This report shows the values of each of the parameters, one scenario per row. Note that the actual alpha value is blank for sample sizes greater than 300, which was the limit set for exact computation.

Most of the report columns have obvious interpretations. Those that may not be obvious are presented here.

### Prop Grp 2 or Control P2

This is the value of P2, the proportion responding positively in the control group.

### Prop|H0 Grp 1 or Trtmnt P1.0

This is the value of P1.0, the proportion responding positively in the treatment group as specified by the null hypothesis. The difference between this value and P2 is the value specified by the null hypothesis.

## Prop|H1 Grp 1 or Trtmnt P1.1

This is the value of P1.1, the proportion responding positively in the treatment group as specified by the alternative hypothesis. The difference between this value and P2 is the value specified by the alternative hypothesis.

## Diff if H0 D0

This is the value of D0, the difference between proportions under the null hypothesis.

## Diff if H1 D1

This is the value of D1, the difference between proportions under the alternative hypothesis.

## Target Alpha

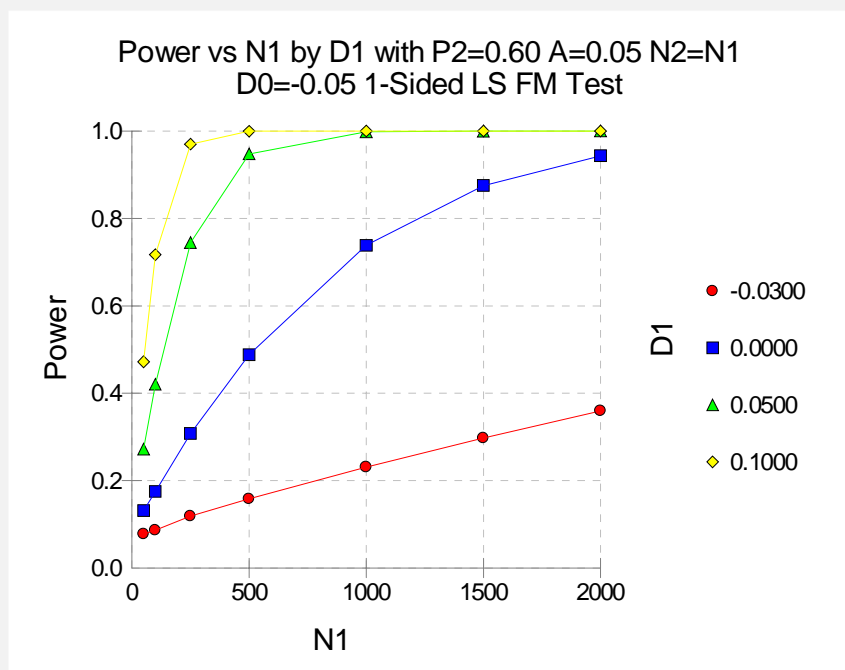
This is the value of alpha that was targeted by the design. Note that the target alpha is not usually achieved exactly.

## Actual Alpha

This is the value of alpha that was actually achieved by this design. Note that since the limit on exact calculations was set to 300, and since this value is calculated exactly, it is not shown for values of N1 greater than 300.

The difference between the Target Alpha and the Actual Alpha is caused by the discrete nature of the binomial distribution and the use of the normal approximation to the binomial in determining the critical value of the test statistic.

## Plots Section



The values from the table are displayed in the above chart. This chart gives us a quick look at the sample size that will be required for various values of D1.



## Example2 - Finding the Sample Size

Continuing with the scenario given in Example 1, the researchers want to determine the sample size needed to achieve 80% power for each value of D1. To cut down on the runtime, they decide to look at approximate values whenever N1 is greater than 100.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.05</b>
D1.....	<b>-0.03 0.00 0.05 0.10</b>
P2.....	<b>0.6</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
N1.....	<i>Ignored since this is the Find setting</i>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100</b>

### Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 – P2**  
**H0: P1-P2<=D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 or Control	Prop H0 Grp 1 or Trtmnt	Prop H1 Grp 1 or Trtmnt	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>	<b>P2</b>	<b>P1.0</b>	<b>P1.1</b>					
0.8000	7491	7491	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0500		0.2000
0.8002	1186	1186	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500		0.1998
0.8008	290	290	0.6000	0.5500	0.6500	-0.0500	0.0500	0.0500		0.1992
0.8011	125	125	0.6000	0.5500	0.7000	-0.0500	0.1000	0.0500		0.1989

The required sample size will depend a great deal on the value of D1. The researchers should spend time determining the most accurate value for D1.

## Example3 – Comparing the Power of Several Test Statistics

Continuing with Example 1, the researchers want to determine which of the eight possible test statistics to adopt by using the comparative reports and charts that *PASS* produces. They decide to compare the powers and actual alphas for various sample sizes between 50 and 200 when D1 is 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>-0.05</b>
D1 .....	<b>0.05</b>
P2.....	<b>0.6</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
N1 .....	<b>50 100 150 200</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>300</b>
<b>Reports Tab</b>	
Show Numeric Report.....	<b>Not checked</b>
Show Comparative Reports .....	<b>Checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots.....	<b>Checked</b>
Summary Statement Rows .....	<b>0</b>

### Annotated Output

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

## Numeric Results

### Power Comparison of Tests Based on the Difference: $P_1 - P_2$

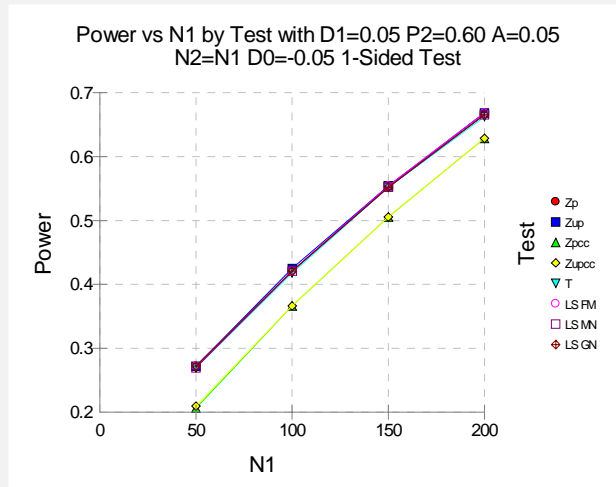
$H_0: P_1 - P_2 \leq D_0$ .  $H_1: P_1 - P_2 = D_1 > D_0$ .

N1/N2	P2	P1.1	Target Alpha	Z(P) Test Power	Z(UnP) Test Power	Z(P) CC Test Power	Z(UnP) CC Test Power	T Test Power	F.M. Score Power	M.N. Score Power	G.N. Score Power
50/50	0.6000	0.6500	0.0500	0.2720	0.2720	0.2064	0.2096	0.2694	0.2720	0.2694	0.2720
100/100	0.6000	0.6500	0.0500	0.4207	0.4248	0.3663	0.3663	0.4178	0.4207	0.4207	0.4207
150/150	0.6000	0.6500	0.0500	0.5540	0.5540	0.5054	0.5054	0.5519	0.5540	0.5519	0.5519
200/200	0.6000	0.6500	0.0500	0.6654	0.6683	0.6286	0.6286	0.6624	0.6683	0.6654	0.6654

### Actual Alpha Comparison of Tests Based on the Difference: $P_1 - P_2$

$H_0: P_1 - P_2 \leq D_0$ .  $H_1: P_1 - P_2 = D_1 > D_0$ .

N1/N2	P2	P1.1	Target Alpha	Z(P) Test Alpha	Z(UnP) Test Alpha	Z(P) CC Test Alpha	Z(UnP) CC Test Alpha	T Test Alpha	F.M. Score Alpha	M.N. Score Alpha	G.N. Score Alpha
50/50	0.6000	0.6500	0.0500	0.0527	0.0527	0.0342	0.0343	0.0526	0.0527	0.0526	0.0527
100/100	0.6000	0.6500	0.0500	0.0499	0.0500	0.0369	0.0369	0.0499	0.0499	0.0499	0.0499
150/150	0.6000	0.6500	0.0500	0.0509	0.0509	0.0398	0.0398	0.0509	0.0509	0.0509	0.0509
200/200	0.6000	0.6500	0.0500	0.0479	0.0482	0.0387	0.0387	0.0477	0.0482	0.0479	0.0479



It is interesting to note that the powers of the continuity-corrected test statistics are consistently lower than the other tests. This occurs because the actual alpha achieved by these tests is lower than for the other tests.

## Example4 - Validation using Machin et al. with Equal Sample Sizes

Machin et al. (1997), page 106, present a sample size study in which  $P_2 = 0.5$ ,  $D_0 = -0.2$ ,  $D_1 = 0$ , one-sided  $\alpha = 0.1$ , and  $\beta = 0.2$ . Using the Farrington and Manning test statistic, they found the sample size to be 55 in each group.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>-0.2</b>
D1 .....	<b>0</b>
P2 .....	<b>0.50</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
N1 .....	<i>Ignored since this is the Find setting</i>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha .....	<b>0.10</b>
Beta .....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Tests Based on the Difference:  $P_1 - P_2$   
H0:  $P_1 - P_2 \leq D_0$ . H1:  $P_1 - P_2 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 or Control	Prop H0 Grp 1 or Trtmnt	Prop H1 Grp 1 or Trtmnt	Diff if H0	Diff if H1	Target Alpha	Actual Alpha	Beta
Power	N1	N2	P2	P1.0	P1.1	D0	D1			
0.8001	55	55	0.5000	0.3000	0.5000	-0.2000	0.0000	0.1000		0.1999

*PASS* found the required sample size to be 55, which corresponds to the results of Machin et al.

## Example5 - Validation using Farrington and Manning

Farrington and Manning (1990), page 1451, present a sample size study in which  $P_2 = 0.05$ ,  $D_0 = 0.2$ ,  $D_1 = 0.35$ , one-sided  $\alpha = 0.05$ , and  $\beta = 0.20$ . Using the Farrington and Manning test statistic, they found the sample size to be 80 in each group. They mention that the true power is 0.813.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>0.2</b>
D1.....	<b>0.35</b>
P2.....	<b>0.05</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
N1.....	<i>Ignored since this is the Find setting</i>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Tests Based on the Difference:  $P_1 - P_2$   
 H0:  $P_1 - P_2 \leq D_0$ . H1:  $P_1 - P_2 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
Power	N1	N2								
0.8007	80	80	0.0500	0.2500	0.4000	0.2000	0.3500	0.0500		0.1993

*PASS* also calculated the required sample size to be 80.

Next, to calculate the exact power for this sample size, we make the following changes to the template.

### Data Tab

Find ..... **Beta and Power**

N1 ..... **80**

### Options Tab

Maximum N1 or N2 Exact ..... **300 (Set >80 to force exact calculation.)**

## Numeric Results

**Numeric Results of Tests Based on the Difference:  $P_1 - P_2$**   
**H0:  $P_2 - P_1 \leq D_0$ . H1:  $P_2 - P_1 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)**

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>								
0.8132	80	80	0.0500	0.2500	0.4000	0.2000	0.3500	0.0500	0.0553	0.1868

*PASS* also calculated the exact power to be 0.813.

## Example6 - Validation of Risk Ratio Calculations using Blackwelder

Blackwelder (1993), page 695, presents a table of power values for several scenarios using the risk ratio. The second line of the table presents the results for the following scenario:  $P_2 = 0.04$ ,  $R_0 = 0.3$ ,  $R_1 = 0.1$ ,  $N_1 = N_2 = 1044$ , one-sided alpha = 0.05, and beta = 0.20. Using the Farrington and Manning likelihood-score test statistic, he found the exact power to be 0.812, the exact alpha to be 0.044, and, using the asymptotic formula, the approximate power to be 0.794.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
R0.....	<b>0.3</b>
R1.....	<b>0.1</b>
P2.....	<b>0.04</b>
Alternative Hypothesis .....	<b>One-Sided (H1:R1&lt;R0)</b>
N1.....	<b>1044</b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2000 (Set high for exact results.)</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Tests Based on the Ratio: P1 / P2  
H0: P1/P2>=R0. H1: P1/P2=R1<R0. Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Ratio if H0 R0	Ratio if H1 R1	Target Alpha	Actual Alpha	Beta
Power	1044	1044	0.0400	0.0120	0.0040	0.300	0.100	0.0500	0.0444	0.1882

PASS also calculated the exact power to be 0.812 and the actual alpha to be 0.044, after rounding.

Next, to calculate the asymptotic power, we make the following changes to the template.

### Options Tab

Maximum N1 or N2 Exact ..... **2 (Set < 1044 to force asymptotic calculation.)**

### Numeric Results

Numeric Results of Tests Based on the Difference:  $P_1 - P_2$   
 $H_0: P_1/P_2 \geq R_0$ .  $H_1: P_1/P_2 = R_1 < R_0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Ratio if H0 R0	Ratio if H1 R1	Target Alpha	Actual Alpha	Beta
Power	1044	1044	0.0400	0.0120	0.0040	0.300	0.100	0.0500		0.2063

*PASS* also calculated the asymptotic power to be 0.794.



## Example7 - Finding the Power after Completing an Experiment

Researchers are studying the effectiveness of a new treatment for cancer. Historically, the standard treatment has enjoyed a 52% cure rate. The new experimental treatment is believed to be better, but it costs much more to administer. After weighing cost versus effectiveness, the researchers decided that they will adopt the new treatment if the cure rate is at least 59%. They conduct a study in which 200 patients are given the new treatment, and 200 are given the standard regimen. They find that 66% are cured by the new treatment, while 52% are cured by the standard treatment. The Farrington and Manning likelihood score test, however, indicates that the results are not statistically significant for  $\alpha = 0.05$ . They now desire to compute the power for a range of alternative values.

Note that a range of alternatives is used in computing the power instead of the actual difference from the study. The power should be computed at values representing practically significant differences from the null value.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta &amp; Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>0.07</b>
D1.....	<b>0.08 to 0.20 by 0.02</b>
P2.....	<b>0.52</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
N1.....	<b>200</b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>200</b>

## Annotated Output

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

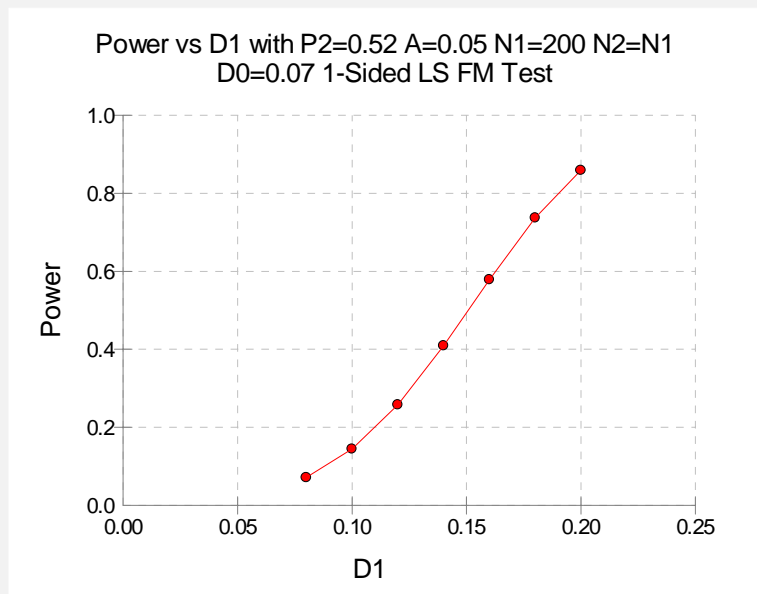
### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 - P2**

**H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Target Alpha	Actual Alpha	Beta
Power										
0.0715	200	200	0.5200	0.5900	0.6000	0.0700	0.0800	0.0500	0.0479	0.9285
0.1446	200	200	0.5200	0.5900	0.6200	0.0700	0.1000	0.0500	0.0479	0.8554
0.2581	200	200	0.5200	0.5900	0.6400	0.0700	0.1200	0.0500	0.0479	0.7419
0.4089	200	200	0.5200	0.5900	0.6600	0.0700	0.1400	0.0500	0.0479	0.5911
0.5783	200	200	0.5200	0.5900	0.6800	0.0700	0.1600	0.0500	0.0479	0.4217
0.7368	200	200	0.5200	0.5900	0.7000	0.0700	0.1800	0.0500	0.0479	0.2632
0.8591	200	200	0.5200	0.5900	0.7200	0.0700	0.2000	0.0500	0.0479	0.1409

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 200.



The power depends a great deal on the value of D1 for this sample size. It is evident that the power is quite low for the majority of alternative values studied.

## Example8 - Finding the Sample Size using Ratios

A study is being designed to determine the effectiveness of a new treatment. Researchers would like to know how large of a sample is needed for comparison of the two treatments. The standard treatment has a success rate of 65%. The researchers will adopt the new treatment, which has fewer side effects, if the success rate is at least 90% of the rate for the standard treatment, i.e.  $P1 = 0.9 \times P2$  or  $P1/P2 = 0.9$ . They would like to calculate the sample sizes necessary to achieve 80%, 85%, 90%, and 95% power for the case where the true ratio between the two proportions is 1.1 and  $\alpha = 0.05$ .

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
R0.....	<b>0.9</b>
R1.....	<b>1.1</b>
P2.....	<b>0.65</b>
Alternative Hypothesis .....	<b>One-Sided (H1:R1&gt;R0)</b>
N1.....	<i>Ignored since this is the Find setting</i>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.05 0.10 0.15 0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>300</b>

## Annotated Output

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

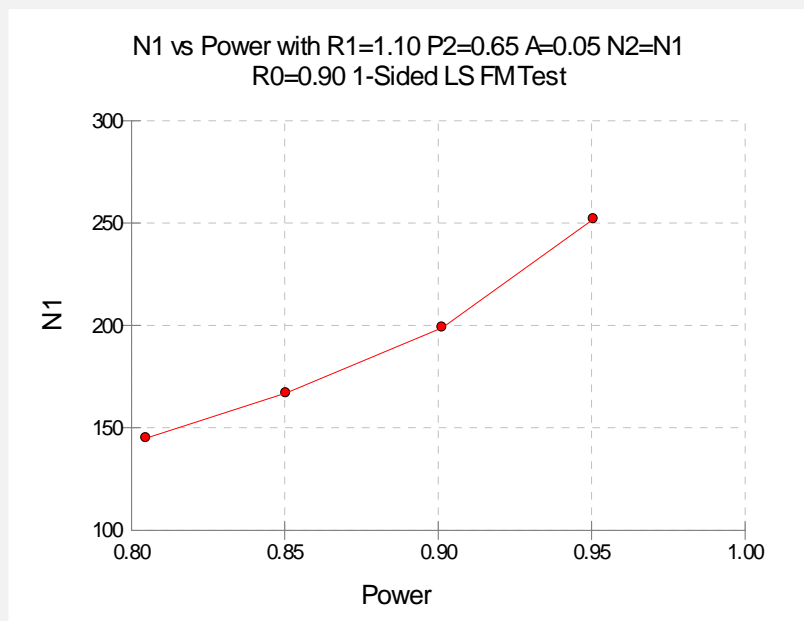
### Numeric Results

Numeric Results of Tests Based on the Ratio:  $P1 / P2$

$H0: P1/P2 \leq R0$ .  $H1: P1/P2 = R1 > R0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 or Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Ratio if H0 R0	Ratio if H1 R1	Target Alpha	Actual Alpha	Beta
Power										
0.9506	252	252	0.6500	0.5850	0.7150	0.900	1.100	0.0500	0.0503	0.0494
0.9013	199	199	0.6500	0.5850	0.7150	0.900	1.100	0.0500	0.0508	0.0987
0.8504	167	167	0.6500	0.5850	0.7150	0.900	1.100	0.0500	0.0507	0.1496
0.8048	145	145	0.6500	0.5850	0.7150	0.900	1.100	0.0500	0.0499	0.1952

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 300.



Necessary sample sizes range from 145 for 80% power to 252 for 95% power for detecting a ratio of 1.1.

## Chapter 210

# Non-Inferiority of Two Independent Proportions

## Introduction

This module provides power analysis and sample size calculation for non-inferiority and superiority tests in two-sample designs in which the outcome is binary. Users may choose from among eight popular test statistics commonly used for running the hypothesis test.

The power calculations assume that independent, random samples are drawn from two populations.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Example

A non-inferiority test example will set the stage for the discussion of the terminology that follows. Suppose that the current treatment for a disease works 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the first questions that must be answered is whether the new treatment is as good as the current treatment. In other words, do at least 70% of treated subjects respond to the new treatment?

Because of the many benefits of the new treatment, clinicians are willing to adopt the new treatment even if it is slightly less effective than the current treatment. They must determine, however, how much less effective the new treatment can be and still be adopted. Should it be adopted if 69% respond? 68%? 65%? 60%? There is a percentage below 70% at which the difference between the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, it was decided that if a response of at least 63% were achieved, the new treatment would be adopted. The difference between these two percentages is called the *margin of equivalence*. The margin of equivalence in this example is 7%.

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is at least 0.63. The statistical hypothesis to be tested is

$$H_0: p_1 - p_2 \leq -0.07 \text{ versus } H_1: p_1 - p_2 > -0.07$$

Notice that when the null hypothesis is rejected, the conclusion is that the response rate is at least 0.63. Note that even though the response rate of the current treatment is 0.70, the hypothesis test is about a response rate of 0.63. Also notice that a rejection of the null hypothesis results in the conclusion of interest.

## Technical Details

The details of sample size calculation for the two-sample design for binary outcomes are presented in the chapter “Two Proportions Non-Null Case,” and they will not be duplicated here. Instead, this chapter only discusses those changes necessary for non-inferiority and superiority tests.

Approximate sample size formulas for non-inferiority tests of two proportions are presented in Chow et al. (2003), page 90. Only large sample (normal approximation) results are given there. The results available in this module use exact calculations based on the enumeration of all possible values in the binomial distribution.

Suppose you have two populations from which dichotomous (binary) responses will be recorded. Assume without loss of generality that the higher proportions are better. The probability (or risk) of cure in population 1 (the treatment group) is  $p_1$  and in population 2 (the reference group) is  $p_2$ . Random samples of  $n_1$  and  $n_2$  individuals are obtained from these two populations. The data from these samples can be displayed in a 2-by-2 contingency table as follows

Group	Success	Failure	Total
Treatment	$x_{11}$	$x_{12}$	$n_1$
Control	$x_{21}$	$x_{22}$	$n_2$
Totals	$m_1$	$m_2$	$N$

The binomial proportions,  $p_1$  and  $p_2$ , are estimated from these data using the formulae

$$\hat{p}_1 = \frac{a}{m} = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{b}{n} = \frac{x_{21}}{n_2}$$

Let  $p_{1,0}$  represent the group 1 proportion tested by the null hypothesis,  $H_0$ . The power of a test is computed at a specific value of the proportion which we will call  $p_{1,1}$ . Let  $\delta$  represent the smallest difference (margin of equivalence) between the two proportions that still results in the conclusion that the new treatment is not inferior to the current treatment. For a non-inferiority test,  $\delta < 0$ . The set of statistical hypotheses that are tested is

$$H_0: p_{1,0} - p_2 \leq \delta \text{ versus } H_1: p_{1,0} - p_2 > \delta$$

which can be rearranged to give

$$H_0: p_{1,0} \leq p_2 + \delta \text{ versus } H_1: p_{1,0} > p_2 + \delta$$

There are three common methods of specifying the margin of equivalence. The most direct is to simply give values for  $p_2$  and  $p_{1,0}$ . However, it is often more meaningful to give  $p_2$  and then specify  $p_{1,0}$  implicitly by specifying the difference, ratio, or odds ratio. Mathematically, the definitions of these parameterizations are

<u>Parameter</u>	<u>Computation</u>	<u>Hypotheses</u>
Difference	$\delta = p_{1,0} - p_2$	$H_0: p_{1,0} - p_2 \leq \delta_0 \text{ vs. } H_1: p_{1,0} - p_2 > \delta_0, \quad \delta_0 < 0$
Ratio	$\phi = p_{1,0} / p_2$	$H_0: p_1 / p_2 \leq \phi_0 \text{ vs. } H_1: p_1 / p_2 > \phi_0, \quad \phi_0 < 1$
Odds Ratio	$\psi = Odds_{1,0} / Odds_2$	$H_0: o_{1,0} / o_2 \leq \psi_0 \text{ versus } H_1: o_{1,0} / o_2 > \psi_0, \quad \psi_0 < 1$

## Difference

The difference is perhaps the most direct method of comparison between two proportions. It is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One difficulty arises when the event of interest is rare. If a difference of 0.001 occurs when the baseline probability is 0.40, it would be dismissed as being trivial. However, if the baseline probability of a disease is 0.002, a 0.001 decrease would represent a reduction of 50%. Thus interpretation of the difference depends on the baseline probability of the event.

Note that if  $\delta < 0$ , the procedure is called a *non-inferiority test* while if  $\delta > 0$  the procedure is called a *superiority test*.

## Non-Inferiority using a Difference

The following example might help you understand the concept of a *non-inferiority test*. Suppose 60% of patients respond to the current treatment method ( $p_2 = 0.60$ ). If the response rate of the new treatment is no less than 5 percentage points worse ( $\delta = -0.05$ ) than the existing treatment, it will be considered to be noninferior. Substituting these figures into the statistical hypotheses gives

$$H_0: \delta \leq -0.05 \text{ versus } H_1: \delta > -0.05$$

Using the relationship

$$p_{1.0} = p_2 + \delta$$

gives

$$H_0: p_{1.0} \leq 0.55 \text{ versus } H_1: p_{1.0} > 0.55$$

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 55%, which means that the new treatment is not inferior to the current treatment.

## Superiority using a Difference

The following example is intended to help you understand the concept of a *superiority test*. Suppose 60% of patients respond to the current treatment method ( $p_2 = 0.60$ ). If the response rate of the new treatment is at least 10 percentage points better ( $\delta = 0.10$ ), it will be considered to be superior to the existing treatment. Substituting these figures into the statistical hypotheses gives

$$H_0: \delta \leq 0.10 \text{ versus } H_1: \delta > 0.10$$

Using the relationship

$$p_{1.0} = p_2 + \delta$$

gives

$$H_0: p_{1.0} \leq 0.70 \text{ versus } H_1: p_{1.0} > 0.70$$

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 0.70. That is, the conclusion of superiority is that the new treatment's response rate is at least 0.10 more than that of the existing treatment.



## Ratio

The ratio,  $\phi = p_{1.0} / p_2$ , gives the relative change in the probability of the response. Testing non-inferiority and superiority use the formulation

$$H_0: p_{1.0} / p_2 \leq \phi_0 \text{ versus } H_1: p_{1.0} / p_2 > \phi_0$$

The only subtlety is that for non-inferiority tests  $\phi_0 < 1$ , while for superiority tests  $\phi_0 > 1$ .

## Non-Inferiority using a Ratio

The following example might help you understand the concept of *non-inferiority* as defined by the ratio. Suppose that 60% of patients ( $p_2 = 0.60$ ) respond to the current treatment method. If a new treatment decreases the response rate by no more than 10% ( $\phi_0 = 0.90$ ), it will be considered to be noninferior to the standard treatment. Substituting these figures into the statistical hypotheses gives

$$H_0: \phi \leq 0.90 \text{ versus } H_1: \phi > 0.90$$

Using the relationship

$$p_{1.0} = \phi_0 p_2$$

gives

$$H_0: p_{1.0} \leq 0.54 \text{ versus } H_1: p_{1.0} > 0.54$$

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is at least 54%. That is, the conclusion of non-inferiority is that the new treatment's response rate is no worse than 10% less than that of the standard treatment.

## Odds Ratio

The odds ratio,  $\psi = (p_{1.0} / (1 - p_{1.0})) / (p_2 / (1 - p_2))$ , gives the relative change in the odds of the response. Testing non-inferiority and superiority use the same formulation

$$H_0: \psi \leq \psi_0 \text{ versus } H_1: \psi > \psi_0$$

The only difference is that for non-inferiority tests  $\psi_0 < 1$ , while for superiority tests  $\psi_0 > 1$ .

## A Note on Setting the Significance Level, Alpha

Setting the significance level has always been somewhat arbitrary. For planning purposes, the standard has become to set alpha to 0.05 for two-sided tests. Almost universally, when someone states that a result is statistically significant, they mean statistically significant at the 0.05 level.

Although 0.05 may be the standard for two-sided tests, it is not always the standard for one-sided tests, such as non-inferiority tests. Statisticians often recommend that the alpha level for one-sided tests be set at 0.025 since this is the amount put in each tail of a two-sided test.

## Power Calculation

The power for a test statistic that is based on the normal approximation can be computed exactly using two binomial distributions. The following steps are taken to compute the power of these tests.

1. Find the critical value using the standard normal distribution. The critical value,  $z_{critical}$ , is that value of  $z$  that leaves exactly the target value of  $\alpha$  in the appropriate tail of the normal distribution.
2. Compute the value of the test statistic,  $z_t$ , for every combination of  $x_{11}$  and  $x_{21}$ . Note that  $x_{11}$  ranges from 0 to  $n_1$ , and  $x_{21}$  ranges from 0 to  $n_2$ . A small value (around 0.0001) can be added to the zero-cell counts to avoid numerical problems that occur when the cell value is zero.
3. If  $z_t > z_{critical}$ , the combination is in the rejection region. Call all combinations of  $x_{11}$  and  $x_{21}$  that lead to a rejection the set  $A$ .
4. Compute the power for given values of  $p_{1,1}$  and  $p_2$  as

$$1 - \beta = \sum_A \binom{n_1}{x_{11}} p_{1,1}^{x_{11}} q_{1,1}^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_2^{x_{21}} q_2^{n_2 - x_{21}}$$

5. Compute the actual value of  $\alpha$  achieved by the design by substituting  $p_2$  for  $p_{1,1}$  to obtain

$$\alpha^* = \sum_A \binom{n_1}{x_{11}} \binom{n_2}{x_{21}} p_2^{x_{11} + x_{21}} q_2^{n_1 + n_2 - x_{11} - x_{21}}$$

## Asymptotic Approximations

When the values of  $n_1$  and  $n_2$  are large (say over 200), these formulas often take a long time to evaluate. In this case, a large sample approximation can be used. The large sample approximation is made by replacing the values of  $\hat{p}_1$  and  $\hat{p}_2$  in the  $z$  statistic with the corresponding values of  $p_{1,1}$  and  $p_2$ , and then computing the results based on the normal distribution. Note that in large samples, the Farrington and Manning statistic is substituted for the Gart and Nam statistic. Also, for large samples, the results for the odds ratio have not (to our knowledge) been published. In this case, we substitute the calculations based on the ratio formulation.

## Test Statistics

Several test statistics have been proposed for testing whether the difference, ratio, or odds ratio are different from a specified value. The main difference among the several test statistics is in the formula used to compute the standard error used in the denominator. These tests are based on the following z-test

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - c}{\hat{\sigma}}$$

The constant,  $c$ , represents a continuity correction that is applied in some cases. When the continuity correction is not used,  $c$  is zero. In power calculations, the values of  $\hat{p}_1$  and  $\hat{p}_2$  are not known. The corresponding values of  $p_{1.1}$  and  $p_2$  may be reasonable substitutes.

Following is a list of the test statistics available in **PASS**. The availability of several test statistics begs the question of which test statistic one should use. The answer is simple: one should use the test statistic that will be used to analyze the data. You may choose a method because it is a standard in your industry, because it seems to have better statistical properties, or because your statistical package calculates it. Whatever your reasons for selecting a certain test statistic, you should use the same test statistic when doing the analysis after the data have been collected.

### Z Test (Pooled)

This test was first proposed by Karl Pearson in 1900. Although this test is usually expressed directly as a chi-square statistic, it is expressed here as a  $z$  statistic so that it can be more easily used for one-sided hypothesis testing. The proportions are pooled (averaged) in computing the standard error. The formula for the test statistic is

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_1}$$

where

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

### Z Test (Unpooled)

This test statistic does not pool the two proportions in computing the standard error.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_2}$$

where

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

### Z Test with Continuity Correction (Pooled)

This test is the same as Z Test (Pooled), except that a continuity correction is used. Remember that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 + \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_1}$$

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

where  $F$  is -1 for lower-tailed hypotheses and 1 for upper-tailed hypotheses.

### Z Test with Continuity Correction (Unpooled)

This test is the same as the Z Test (Unpooled), except that a continuity correction is used. Remember that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_2}$$

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

where  $F$  is -1 for lower-tailed hypotheses and 1 for upper-tailed hypotheses.

### T-Test of Difference

Because of a detailed, comparative study of the behavior of several tests, D'Agostino (1988) and Upton (1982) proposed using the usual two-sample  $t$ -test for testing whether the two proportions are equal. One substitutes a '1' for a success and a '0' for a failure in the usual, two-sample  $t$ -test formula.

## Miettinen and Nurminen's Likelihood Score Test of the Difference

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the difference is equal to a specified, non-zero, value,  $\delta_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic. The formula for computing this test statistic is

$$z_{MND} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_{MND}}$$

where

$$\hat{\sigma}_{MND} = \sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right) \left( \frac{N}{N-1} \right)}$$

$$\tilde{p}_1 = \tilde{p}_2 + \delta_0$$

$$\tilde{p}_1 = 2B \cos(A) - \frac{L_2}{3L_3}$$

$$A = \frac{1}{3} \left[ \pi + \cos^{-1} \left( \frac{C}{B^3} \right) \right]$$

$$B = \text{sign}(C) \sqrt{\frac{L_2^2}{9L_3} - \frac{L_1}{3L_3}}$$

$$C = \frac{L_2^3}{27L_3^3} - \frac{L_1 L_2}{6L_3^2} + \frac{L_0}{2L_3}$$

$$L_0 = x_{21} \delta_0 (1 - \delta_0)$$

$$L_1 = [N_2 \delta_0 - N - 2x_{21}] \delta_0 + M_1$$

$$L_2 = (N + N_2) \delta_0 - N - M_1$$

$$L_3 = N$$

### Miettinen and Nurminen's Likelihood Score Test of the Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the ratio is equal to a specified value  $\phi_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 / \tilde{p}_2 = \phi_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{MNR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \phi_0^2 \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right) \left( \frac{N}{N-1} \right)}}$$

where

$$\begin{aligned}\tilde{p}_1 &= \tilde{p}_2 \phi_0 \\ \tilde{p}_2 &= \frac{-B - \sqrt{B^2 - 4AC}}{2A} \\ A &= N\phi_0 \\ B &= -[N_1\phi_0 + x_{11} + N_2 + x_{21}\phi_0] \\ C &= M_1\end{aligned}$$

## Miettinen and Nurminen's Likelihood Score Test of the Odds Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the odds ratio is equal to a specified value,  $\psi_0$ . Because the approach they used with the difference and ratio does not easily extend to the odds ratio, they used a score statistic approach for the odds ratio. The regular MLE's are  $\hat{p}_1$  and  $\hat{p}_2$ . The constrained MLE's are  $\tilde{p}_1$  and  $\tilde{p}_2$ . These estimates are constrained so that  $\tilde{\psi} = \psi_0$ . A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{MNO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{N_2 \tilde{p}_1 \tilde{q}_1} + \frac{1}{N_2 \tilde{p}_2 \tilde{q}_2} \right) \left( \frac{N}{N-1} \right)}}$$

where

$$\begin{aligned}\tilde{p}_1 &= \frac{\tilde{p}_2 \psi_0}{1 + \tilde{p}_2 (\psi_0 - 1)} \\ \tilde{p}_2 &= \frac{-B + \sqrt{B^2 - 4AC}}{2A} \\ A &= N_2 (\psi_0 - 1) \\ B &= N_1 \psi_0 + N_2 - M_1 (\psi_0 - 1) \\ C &= -M_1\end{aligned}$$

## Farrington and Manning's Likelihood Score Test of the Difference

Farrington and Manning (1990) proposed a test statistic for testing whether the difference is equal to a specified value  $\delta_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ , are used in the denominator. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMD} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

### Farrington and Manning's Likelihood Score Test of the Ratio

Farrington and Manning (1990) proposed a test statistic for testing whether the ratio is equal to a specified value  $\phi_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 / \tilde{p}_2 = \phi_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to increase the variance estimate. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \phi_0^2 \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

### Farrington and Manning's Likelihood Score Test of the Odds Ratio

Farrington and Manning (1990) indicate that the Miettinen and Nurminen statistic may be modified by removing the factor  $N/(N-1)$ .

The formula for computing this test statistic is

$$z_{FMO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{N_2 \tilde{p}_1 \tilde{q}_1} + \frac{1}{N_2 \tilde{p}_2 \tilde{q}_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.



### Gart and Nam's Likelihood Score Test of the Difference

Gart and Nam (1990), page 638, proposed a modification to the Farrington and Manning (1988) difference test that corrects for skewness. Let  $z_{FMD}(\delta)$  stand for the Farrington and Manning difference test statistic described above. The skewness corrected test statistic,  $z_{GND}$ , is the appropriate solution to the quadratic equation

$$(-\tilde{\gamma})z_{GND}^2 + (-1)z_{GND} + (z_{FMD}(\delta) + \tilde{\gamma}) = 0$$

where

$$\tilde{\gamma} = \frac{\tilde{V}^{3/2}(\delta)}{6} \left( \frac{\tilde{p}_1 \tilde{q}_1 (\tilde{q}_1 - \tilde{p}_1)}{n_1^2} - \frac{\tilde{p}_2 \tilde{q}_2 (\tilde{q}_2 - \tilde{p}_2)}{n_2^2} \right)$$

### Gart and Nam's Likelihood Score Test of the Ratio

Gart and Nam (1988), page 329, proposed a modification to the Farrington and Manning (1988) ratio test that corrects for skewness. Let  $z_{FMR}(\phi)$  stand for the Farrington and Manning ratio test statistic described above. The skewness corrected test statistic,  $z_{GNR}$ , is the appropriate solution to the quadratic equation

$$(-\tilde{\varphi})z_{GNR}^2 + (-1)z_{GNR} + (z_{FMR}(\phi) + \tilde{\varphi}) = 0$$

where

$$\tilde{\varphi} = \frac{1}{6\tilde{u}^{3/2}} \left( \frac{\tilde{q}_1 (\tilde{q}_1 - \tilde{p}_1)}{n_1^2 \tilde{p}_1^2} - \frac{\tilde{q}_2 (\tilde{q}_2 - \tilde{p}_2)}{n_2^2 \tilde{p}_2^2} \right)$$

$$\tilde{u} = \frac{\tilde{q}_1}{n_1 \tilde{p}_1} + \frac{\tilde{q}_2}{n_2 \tilde{p}_2}$$

## 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 Procedure Templates.

## Data Tab (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *P1.1*, *Alpha*, *Beta*, *N1*, and *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.

### Test Statistic

Specify which test statistic is used in searching and reporting. Although the pooled *z*-test is commonly shown in elementary statistics books, the likelihood score test is arguably the best choice.

Note that *C.C.* is an abbreviation for *Continuity Correction*. This refers to the adding or subtracting  $1/(2n)$  to (or from) the numerator of the *z*-value to bring the normal approximation closer to the binomial distribution.

## Higher Proportions Are

This option specifies whether proportions represent successes (better) or failures (worse).

### Better (Successes)

When proportions represent successes, higher proportions are better. A noninferior treatment is one whose proportion is at least almost as high as that of the reference group.

For testing non-inferiority, D0 is negative, R0 is less than 1, and OR0 is less than 1. For testing superiority, D0 is positive, R0 is greater than 1, and OR0 is greater than 1.

### Worse (Failures)

When proportions represent failures, lower proportions are better. A noninferior treatment is one whose proportion is at most almost as low as that of the reference group.

For testing non-inferiority, D0 is positive, R0 is greater than 1, and OR0 is greater than 1. For testing superiority, D0 is negative, R0 is less than 1, and OR0 is less than 1.

## N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. 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  $N2 = N1$ .

## 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 unequal proportions when in fact they are not equal.

Values must be between 0 and 1. Because this is a one-sided test, unless you have good reason to do otherwise, the value of 0.025 is recommended. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.025 0.05* or *0.01 to 0.10 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 when in fact it is false.

Values must be between 0 and 1. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions, P1 and P2. P1.0 is the value of the P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated.

### P1.0 (Equivalence Proportion)

This option allows you to specify the value P1.0 directly. This is that value of treatment group's proportion above which the treatment group is considered noninferior to the reference group.

When *Higher Proportions Are* is set to *Better*, the trivial proportion is the smallest value of P1 for which the treatment group is declared noninferior to the reference group. In this case, P1.0 should be less than P2 for non-inferiority tests and greater than P2 for superiority tests. The reverse is the case when *Higher Proportions Are* is set to *Worse*.

Proportions must be between 0 and 1. They cannot take on the values 0 or 1. This value should not be set to exactly the value of P2.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

### P1.1 (Actual Proportion)

This option specifies the value of P1.1 which is the value of the treatment proportion at which the power is to be calculated. Proportions must be between 0 and 1. They cannot take on the values 0 or 1.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $D0$ , and  $D1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### D0 (Equivalence Difference)

This option specifies the trivial difference (often called the *margin of error*) between  $P1.0$  (the value of  $P1$  under  $H0$ ) and  $P2$ . This difference is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = P2 + D0$ .

When *Higher Proportions Are* is set to *Better*, the trivial difference is that amount by which  $P1$  can be less than  $P2$  and still have the treatment group declared noninferior to the reference group. In this case,  $D0$  should be negative for non-inferiority tests and positive for superiority tests.

The reverse is the case when *Higher Proportions Are* is set to *worse*.

You may enter a range of values such as  $-.03$   $-.05$   $-.10$  or  $-.05$  to  $-.01$  by  $.01$ . Differences must be between  $-1$  and  $1$ .  $D0$  cannot take on the values  $-1$ ,  $0$ , or  $1$ .

### D1 (Actual Difference)

This option specifies the actual difference between  $P1.1$  (the actual value of  $P1$ ) and  $P2$ . This is the value of the difference at which the power is calculated. In non-inferiority trials, this difference is often set to  $0$ .

The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group). This difference is used with  $P2$  to calculate the value of  $P1$  using the formula:  $P1.1 = D1 + P2$ .

You may enter a range of values such as  $-.05$   $0$   $.5$  or  $-.05$  to  $.05$  by  $.02$ . Actual differences must be between  $-1$  and  $1$ . They cannot take on the values  $-1$  or  $1$ .

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $R0$ , and  $R1$  are given, the values of  $P1.0$  and  $P1.1$  can be calculated.

### R0 (Equivalence Ratio)

This option specifies the trivial ratio (also called the Relative Margin of Equivalence) between  $P1.0$  and  $P2$ . The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = R0 \times P2$ .

When *Higher Proportions Are* is set to *Better*, the trivial ratio is the relative amount by which  $P1$  can be less than  $P2$  and still have the treatment group declared noninferior to the reference group. In this case,  $R0$  should be less than one for non-inferiority tests and greater than 1 for superiority tests. The reverse is the case when *Higher Proportions Are* is set to *Worse*.

Ratios must be positive.  $R0$  cannot take on the value of 1. You may enter a range of values such as *0.95 .97 .99* or *.91 to .99 by .02*.

### R1 (Actual Ratio)

This option specifies the ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the actual proportion in the treatment group. The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1. This difference is used with  $P2$  to calculate the value of  $P1$  using the formula:  $P1.1 = R1 \times P2$ . In non-inferiority trials, this ratio is often set to 1.

Ratios must be positive. You may enter a range of values such as *0.95 1 1.05* or *0.9 to 1.9 by 0.02*.

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratios,  $O1.1 / O2$  and  $O1.0 / O2$ . Note that the odds are defined as  $O2 = P2 / (1 - P2)$ ,  $O1.0 = P1.0 / (1 - P1.0)$ , etc.  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $OR0$ , and  $OR1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### OR0 (Equivalence Odds Ratio)

This option specifies the trivial odds ratio between  $P1.0$  and  $P2$ . The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis.  $OR0$  is used with  $P2$  to calculate the value of  $P1.0$ .

When *Higher Proportions Are* is set to *Better*, the trivial odds ratio implicitly gives the amount by which  $P1$  can be less than  $P2$  and still have the treatment group declared noninferior to the reference group. In this case,  $OR0$  should be less than 1 for non-inferiority tests and greater than 1 for superiority tests. The reverse is the case when *Higher Proportions Are* is set to *Worse*.

Odds ratios must be positive.  $OR0$  cannot take on the value of 1.

You may enter a range of values such as *0.95 0.97 0.99* or *0.91 to 0.99 by 0.02*.

### OR1 (Actual Odds Ratio)

This option specifies the odds ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the actual proportion in the treatment group. The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1. This value is used with  $P2$  to calculate the value of  $P1$ . In non-inferiority trials, this odds ratio is often set to 1.

Odds ratios must be positive. You may enter a range of values such as *0.95 1 1.05* or *0.9 to 1 by 0.02*.

### P2 (Reference Group Proportion)

Specify the value of  $p_2$ , the reference, baseline, or control group's proportion. The null hypothesis is that the two proportions differ by no more than a specified amount. Since  $P2$  is a proportion, these values must be between 0 and 1.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

## Options Tab

The Options tab contains various limits and options.

### Maximum N1 or N2 Exact

When either N1 or N2 is above this amount, power calculations are based on the normal approximation to the binomial. In this case, the actual value of alpha is not calculated. Currently, for three-gigahertz computers, a value near 200 is reasonable. As computers get faster, this number may be increased.

### Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

### Zero Count Adjustment Method

Zero cell counts cause many calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option ‘Add to zero cells only.’

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts. We have found that adding 0.0001 seems to work well.

### Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This is the amount that is added. We have found that 0.0001 works well.

Be warned that the value of the ratio and the odds ratio will be affected by the amount specified here!



## Example1 - Finding Power

A study is being designed to establish the non-inferiority of a new treatment compared to the current treatment. Historically, the current treatment has enjoyed a 60% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the current treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the current treatment. The researchers will recommend adoption of the new treatment if it has a cure rate of at least 55%.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data that will be (or has been) obtained. They want to study the power of the Farrington and Manning test at group sample sizes ranging from 50 to 500 for detecting a difference of -0.05 when the actual cure rate of the new treatment ranges from 57% to 70%. The significance level will be 0.025.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.05</b>
D1.....	<b>-0.03 0.00 0.05 0.10</b>
P2.....	<b>0.6</b>
Higher Proportions Are.....	<b>Better</b>
N1.....	<b>50 to 500 by 50</b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.025</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>300</b>

### Annotated Output

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

## Numeric Results

**Numeric Results of Non-Inferiority Tests Based on the Difference: P1 – P2**  
**H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Sample Size Grp 1	Sample Size Grp 2	Grp 2 Prop P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>								
0.0380	50	50	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250	0.0236	0.9620
0.0494	100	100	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250	0.0267	0.9506
0.0525	150	150	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250	0.0241	0.9475
0.0588	200	200	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250	0.0244	0.9412
0.0650	250	250	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250	0.0241	0.9350
0.0735	300	300	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250	0.0261	0.9265
0.0776	350	350	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250		0.9224
0.0832	400	400	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250		0.9168
0.0886	450	450	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250		0.9114
0.0940	500	500	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250		0.9060
Report continues ...										

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 300.

### Report Definitions

'Power' is the probability of rejecting a false null hypothesis. It should be close to one.

'N1 and N2' are the sizes of the samples drawn from the corresponding groups.

'P2' is the response rate for group two which is the standard, reference, baseline, or control group.

'P1.0' is the smallest treatment-group response rate that still yields a non-inferiority conclusion.

'P1.1' is the treatment-group response rate at which the power is calculated.

'D0' is the non-inferiority margin. It is the difference P1-P2 assuming H0.

'D1' is the actual difference, P1-P2, at which the power is calculated.

'Target Alpha' is the probability of rejecting a true null hypothesis that was desired.

'Actual Alpha' is the value of alpha that is actually achieved.

'Beta' is the probability of accepting a false H0. Beta = 1 - Power.

'Grp 1' refers to Group 1 which is the treatment or experimental group.

'Grp 2' refers to Group 2 which is the reference, standard, or control group.

'Equiv.' refers to a small amount that is not of practical importance.

'Actual' refers to the true value at which the power is computed.

### Summary Statements

Sample sizes of 50 in group one and 50 in group two achieve 4% power to detect a non-inferiority margin difference between the group proportions of -0.0500. The reference group proportion is 0.6000. The treatment group proportion is assumed to be 0.5500 under the null hypothesis of inferiority. The power was computed at for the case when the actual treatment group proportion is 0.5700. The test statistic used is the one-sided Score test (Farrington & Manning). The significance level of the test was targeted at 0.0250. The significance level actually achieved by this design is 0.0236.

This report shows the values of each of the parameters, one scenario per row. Note that the actual alpha value is blank for sample sizes greater than 300, which was the limit set for exact computation.

Most of the report columns have obvious interpretations. Those that may not be obvious are presented here. Note that the discussion below assumes that higher response rates are better and that non-inferiority testing (rather than superiority testing) is planned.

## Prop Grp 2 P2

This is the value of P2, the response rate in the control group.

## Equiv. Grp 1 Prop P1.0

This is the value of P1.0, the response rate of the treatment group, as specified by the null hypothesis of inferiority. Values of P1 less than this amount are considered different from P2. Values of P1 greater than this are considered noninferior to the reference group. The difference between this value and P2 is the value of the null hypothesis.

## Actual Grp 1 Prop P1.1

This is the value of P1.1, the response rate of the treatment group, at which the power is computed. This is the value of P1 under the alternative hypothesis. The difference between this value and P2 is the value of the alternative hypothesis.

## Equiv. Margin Diff D0

This is the value of D0, the difference between the two group proportions under the null hypothesis. This value is often called the *margin of non-inferiority*.

## Actual Margin Diff D1

This is the value of D1, the difference between the two group proportions at which the power is computed. This is the value of the difference under the alternative hypothesis.

## Target Alpha

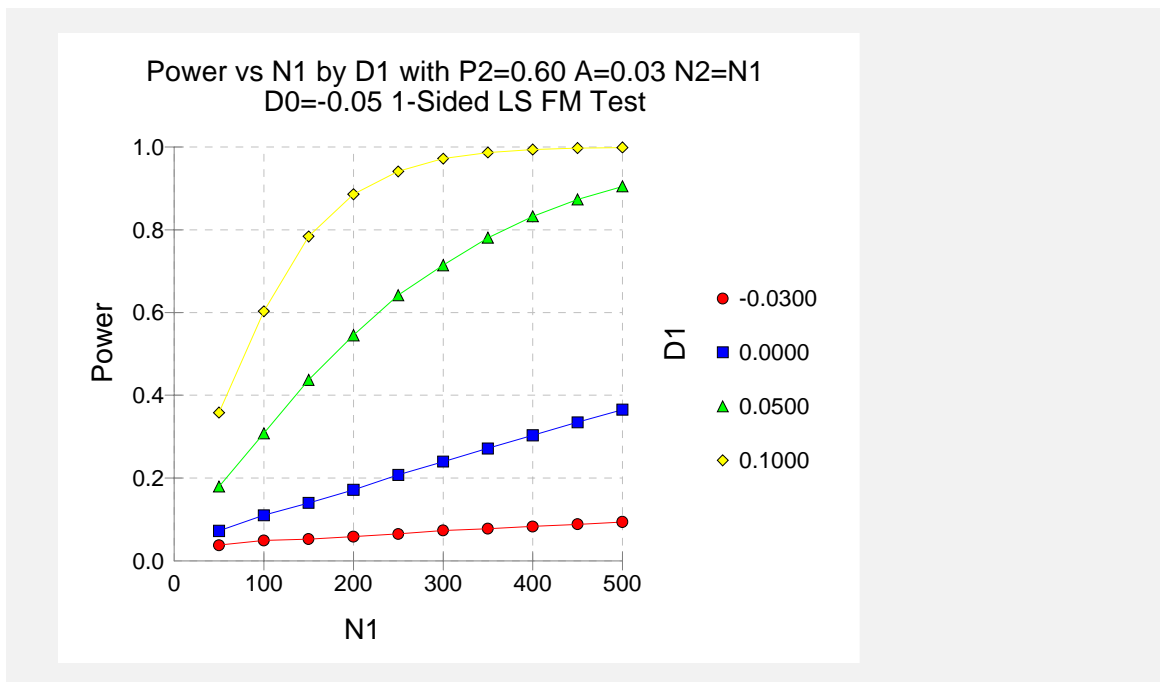
This is the value of alpha that was targeted by the design. Note that the target alpha is not usually achieved exactly. For one-sided tests, this value should usually be 0.025.

## Actual Alpha

This is the value of alpha that was actually achieved by this design. Note that since the limit on exact calculations was set to 300, and since this value is calculated exactly, it is not shown for values of N1 greater than 300.

The difference between the Target Alpha and the Actual Alpha is caused by the discrete nature of the binomial distribution and the use of the normal approximation to the binomial in determining the critical value of the test statistic.

## Plots Section



The values from the table are displayed in the above chart. This chart gives us a quick look at the sample size that will be required for various values of D1.

## Example2 - Finding the Sample Size

Continuing with the scenario given in Example 1, the researchers want to determine the sample size necessary for each value of D1 to achieve a power of 0.80. To cut down on the runtime, they decide to look at approximate values whenever N1 is greater than 100.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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**  
 Test Statistic ..... **Likelihood Score (Farr. & Mann.)**  
 D0 ..... **-0.05**  
 D1 ..... **-0.03 0.00 0.05 0.10**  
 P2 ..... **0.6**  
 Higher Proportions Are ..... **Better**  
 N1 ..... *Ignored since this is the Find setting*  
 N2 ..... **Use R**  
 R ..... **1.0**  
 Alpha ..... **0.025**  
 Beta ..... **0.2**

##### Options Tab

Maximum N1 or N2 Exact ..... **100**

### Annotated Output

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

### Numeric Results

**Numeric Results of Non-Inferiority Tests Based on the Difference: P1 – P2**  
**H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Target Alpha	Actual Alpha	Beta
Power	9509	9509	0.6000	0.5500	0.5700	-0.0500	-0.0300	0.0250		0.2000
0.8000	1505	1505	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0250		0.1999
0.8001	368	368	0.6000	0.5500	0.6500	-0.0500	0.0500	0.0250		0.1992
0.8008	159	159	0.6000	0.5500	0.7000	-0.0500	0.1000	0.0250		0.1981

The required sample size will depend a great deal on the value of D1. Any effort spent determining an accurate value for D1 will be worthwhile.

## Example3 – Comparing the Power of Several Test Statistics

Continuing with Example 1, the researchers want to determine which of the eight possible test statistics to adopt by using the comparative reports and charts that *PASS* produces. They decide to compare the powers and actual alphas for various sample sizes between 50 and 200 when D1 is 0.1.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta &amp; Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.05</b>
D1.....	<b>0.10</b>
P2.....	<b>0.6</b>
Higher Proportions Are.....	<b>Better</b>
N1.....	<b>50 100 150 200</b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.025</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>300</b>
<b>Reports Tab</b>	
Show Numeric Report .....	<b>Not checked</b>
Show Comparative Reports .....	<b>Checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots.....	<b>Checked</b>
Summary Statement Rows.....	<b>0</b>

### Annotated Output

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

## Numeric Results

### Power Comparison of Non-Inferiority Tests Based on the Difference: $P_1 - P_2$

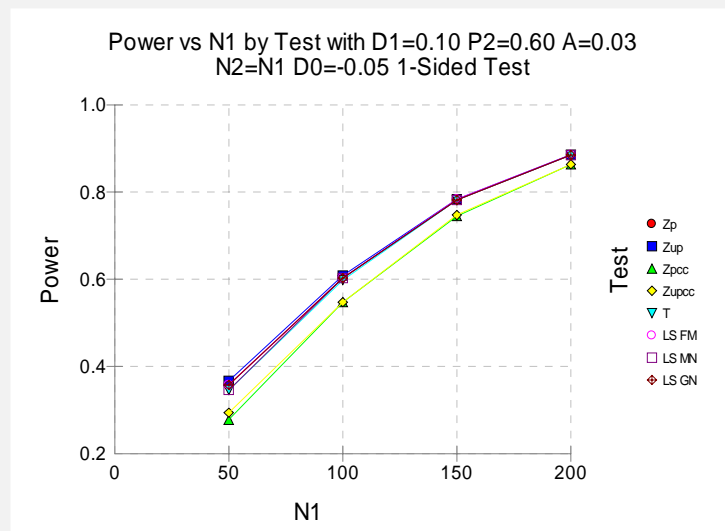
$H_0: P_1 - P_2 \leq D_0$ .  $H_1: P_1 - P_2 = D_1 > D_0$ .

N1/N2	P2	P1	Target Alpha	Z(P) Test Power	Z(UnP) Test Power	Z(P) CC Test Power	Z(UnP) CC Test Power	T Test Power	F.M. Score Power	M.N. Score Power	G.N. Score Power
50/50	0.6000	0.7000	0.0250	0.3581	0.3670	0.2782	0.2945	0.3464	0.3581	0.3464	0.3581
100/100	0.6000	0.7000	0.0250	0.6030	0.6088	0.5474	0.5475	0.5982	0.6030	0.6030	0.6030
150/150	0.6000	0.7000	0.0250	0.7821	0.7837	0.7453	0.7474	0.7821	0.7837	0.7821	0.7821
200/200	0.6000	0.7000	0.0250	0.8849	0.8857	0.8635	0.8638	0.8849	0.8857	0.8849	0.8849

### Actual Alpha Comparison of Non-Inferiority Tests Based on the Difference: $P_1 - P_2$

$H_0: P_2 - P_1 \leq D_0$ .  $H_1: P_2 - P_1 = D_1 > D_0$ .

N1/N2	P1	P2	Target Alpha	Z(P) Test Alpha	Z(UnP) Test Alpha	Z(P) CC Test Alpha	Z(UnP) CC Test Alpha	T Test Alpha	F.M. Score Alpha	M.N. Score Alpha	G.N. Score Alpha
50/50	0.6000	0.7000	0.0250	0.0236	0.0253	0.0140	0.0161	0.0225	0.0236	0.0225	0.0236
100/100	0.6000	0.7000	0.0250	0.0267	0.0267	0.0190	0.0190	0.0266	0.0267	0.0267	0.0267
150/150	0.6000	0.7000	0.0250	0.0239	0.0241	0.0181	0.0183	0.0239	0.0241	0.0239	0.0239
200/200	0.6000	0.7000	0.0250	0.0243	0.0244	0.0191	0.0191	0.0243	0.0244	0.0243	0.0243



It is interesting to note that the powers of the continuity-corrected test statistics are consistently lower than the other tests. This occurs because the actual alpha achieved by these tests is lower than for the other tests. An interesting finding of this example is that the regular  $t$ -test performed about as well as the  $z$ -test.

## Example4 - Validation using Machin with Equal Sample Sizes

Machin et al. (1997), page 106, present a sample size study in which  $P_2 = 0.5$ ,  $D_0 = -0.2$ ,  $D_1 = 0$ , one-sided  $\alpha = 0.1$ , and  $\beta = 0.2$ . Using the Farrington and Manning test statistic, they found the sample size to be 55 in each group.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.2</b>
D1.....	<b>0.0</b>
P2.....	<b>0.5</b>
Higher Proportions Are.....	<b>Better</b>
N1.....	<i>Ignored since this is the Find setting</i>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.1</b>
Beta.....	<b>0.2</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Non-Inferiority Tests Based on the Difference:  $P_1 - P_2$   
 $H_0: P_1 - P_2 \leq D_0$ .  $H_1: P_1 - P_2 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2	Equiv. Grp 1 Prop	Actual Grp 1 Prop	Equiv. Margin Diff	Actual Margin Diff	Target Alpha	Actual Alpha	Beta
Power	N1	N2	P2	P1.0	P1.1	D0	D1			
0.8001	55	55	0.5000	0.3000	0.5000	-0.2000	0.0000	0.1000		0.1999

*PASS* found the required sample size to be 55 which corresponds to Machin.

## Example5 - Validation of a Superiority Test using Farrington and Manning

Farrington and Manning (1990), page 1451, present a sample size study for a superiority test in which  $P_2 = 0.05$ ,  $D_0 = 0.2$ ,  $D_1 = 0.35$ , one-sided  $\alpha = 0.05$ , and  $\beta = 0.20$ . Using the Farrington and Manning test statistic, they found the sample size to be 80 in each group. They mention that the true power is 0.813.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>0.2</b>
D1 .....	<b>0.35</b>
P2.....	<b>0.05</b>
Higher Proportions Are .....	<b>Better</b>
N1 .....	<i>Ignored since this is the search parameter.</i>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Non-Inferiority Tests Based on the Difference:  $P_1 - P_2$   
H0:  $P_1 - P_2 \leq D_0$ . H1:  $P_1 - P_2 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Target Alpha	Actual Alpha	Beta
Power 0.8007	N1 80	N2 80	0.0500	0.2500	0.4000	0.2000	0.3500	0.0500		0.1993

*PASS* also calculated the required sample size to be 80.



Next, to calculate the exact power for this sample size, we make the following changes to the template.

### Data Tab

Find ..... **Beta and Power**

N1 ..... **80**

### Options Tab

Maximum N1 or N2 Exact ..... **200** (Set >80 to force exact calculation.)

## Numeric Results

**Numeric Results of Non-Inferiority Tests Based on the Difference: P1 – P2**  
H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>								
0.8132	80	80	0.0500	0.2500	0.4000	0.2000	0.3500	0.0500	0.0553	0.1993

*PASS* also calculated the exact power to be 0.813.

## Example 6 - Validation of Risk Ratio Calculations using Blackwelder

Blackwelder (1993), page 695, presents a table of power values for several scenarios using the risk ratio. The second line of the table presents the results for the following scenario:  $P_2 = 0.04$ ,  $R_0 = 0.3$ ,  $R_1 = 0.1$ ,  $N_1 = N_2 = 1044$ , one-sided  $\alpha = 0.05$ , and  $\beta = 0.20$ . Using the Farrington and Manning likelihood-score test statistic, he found the exact power to be 0.812, the exact alpha to be 0.044, and, using the asymptotic formula, the approximate power to be 0.794.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the ratio parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
R0 .....	<b>0.3</b>
R1 .....	<b>0.1</b>
P2 .....	<b>0.04</b>
Higher Proportions Are .....	<b>Worse</b>
N1 .....	<b>1044</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2000 (Set high for exact results.)</b>

### Annotated Output

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

### Numeric Results

**Numeric Results of Non-Inferiority Tests Based on the Difference: P1 / P2**  
H0:  $P_1/P_2 \geq R_0$ . H1:  $P_1/P_2 = R_1 < R_0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Ratio R0	Actual Margin Ratio R1	Target Alpha	Actual Alpha	Beta
<b>Power</b>	<b>N1</b>	<b>N2</b>								
0.8118	1044	1044	0.0400	0.0120	0.0040	0.300	0.100	0.0500	0.0444	0.1882

PASS also calculated the power to be 0.812 and the actual alpha to be 0.044, within rounding.

Next, to calculate the asymptotic power, we make the following changes to the template.

### Options Tab

Maximum N1 or N2 Exact .....2 (Set < 1044 to force asymptotic calculation.)

### Numeric Results

Numeric Results of Non-Inferiority Tests Based on the Difference: P1 / P2  
H0:  $P1/P2 \geq R0$ . H1:  $P1/P2 = R1 < R0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Ratio R0	Actual Margin Ratio R1	Target Alpha	Actual Alpha	Beta
Power	1044	1044	0.0400	0.0120	0.0040	0.300	0.100	0.0500		0.2063

*PASS* also calculated the power to be 0.794.

## Example7 - Finding Power following an Experiment

In an effort to show a new treatment non-inferior to the current standard, researchers randomly assigned 80 subjects to each treatment. The new treatment was to be considered non-inferior if the odds ratio (treatment to standard) was at least 0.80. Using the Farrington and Manning Likelihood Score test, non-inferiority could not be concluded. The researchers now want to see the power of the test. The control proportion was 0.625.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
Specify Treatment Proportion using..	<b>Odds Ratios</b>
OR0.....	<b>0.80</b>
OR1.....	<b>1.0</b>
P2.....	<b>0.625</b>
Higher Proportions Are .....	<b>Better</b>
N1 .....	<b>80</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored</i>

### Annotated Output

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

### Numeric Results

**Numeric Results for Non-Inferiority Tests Based on the Odds Ratio: O1 / O2**  
H0:  $O1/O2 \leq OR0$ . H1:  $O1/O2 = OR1 > OR0$ . Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Grp 2 Prop P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin O.R. OR0	Actual Margin O.R. OR1	Target Alpha	Actual Alpha	Beta
Power	N1	N2								
0.1845	80	80	0.6250	0.5714	0.6250	0.800	1.000	0.0500	0.0571	0.8155

The power of a test with 80 receiving each treatment is only 0.1801.

## Example8 - Finding True Proportion Difference

Researchers have developed a new treatment with minimal side effects compared to the standard treatment. The researchers are limited by the number of subjects (140 per group) they can use to show the new treatment is non-inferior. The new treatment will be deemed non-inferior if it is at least 0.10 below the success rate of the standard treatment. The standard treatment has a success rate of about 0.75. The researchers want to know how much more successful the new treatment must be (in truth) to yield a test which has 90% power. The test statistic used will be the pooled Z test.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>P1.1 (Search&gt;P1.0)</b>
Test Statistic.....	<b>Z Test (Pooled)</b>
D0.....	<b>-0.10</b>
D1.....	<i>Ignored</i>
P2.....	<b>0.75</b>
Higher Proportions Are.....	<b>Better</b>
N1.....	<b>140</b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.10</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>500 (Set high for exact results.)</b>

## Annotated Output

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

### Numeric Results

**Numeric Results for Non-Inferiority Tests Based on the Difference: P1 - P2**  
H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Z test (pooled)

	Sample Size Grp 1	Sample Size Grp 2	Grp 2 Prop P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Target Alpha	Actual Alpha	Beta
Power	N1	N2								
0.9000	140	140	0.7500	0.6500	0.7961	-0.1000	0.0461	0.0500	0.0505	0.1000

With 140 subjects in each group, the new treatment must have a success rate 0.0464 higher than the current treatment (or about 0.7964) to have 90% power in the test of non-inferiority.

## Chapter 215

# Equivalence of Two Independent Proportions

## Introduction

This module provides power analysis and sample size calculation for equivalence tests in two-sample designs in which the outcome is binary. Users may choose from among eight popular test statistics commonly used for running the hypothesis test.

The power calculations assume that independent, random samples are drawn from two populations.

## Four Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, ratios of proportions, and odds ratios. Each of these options is listed separately on the menus.

## Example

An equivalence test example will set the stage for the discussion of the terminology that follows. Suppose that the response rate of the standard treatment of a disease is 0.70. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the first questions that must be answered is whether the new treatment is therapeutically equivalent to the standard treatment.

Because of the many benefits of the new treatment, clinicians are willing to adopt the new treatment even if its effectiveness is slightly different from the standard. After thoughtful discussion with several clinicians, it is decided that if the response rate of the new treatment is between 0.63 and 0.77, the new treatment would be adopted. The *margin of equivalence* is 0.07.

The developers must design an experiment to test the hypothesis that the response rate of the new treatment does not differ from that of the standard treatment by more than 0.07. The statistical hypothesis to be tested is

$$H_0: |p_1 - p_2| \geq 0.07 \text{ versus } H_1: |p_1 - p_2| < 0.07$$

## Technical Details

The details of sample size calculation for the two-sample design for binary outcomes are presented in the chapter entitled “Two Proportion Non-Null Case,” and they will not be duplicated here. Instead, this chapter only discusses those changes necessary for equivalence tests.

Approximate sample size formulas for equivalence tests of two proportions are presented in Chow et al. (2003), page 88. Only large sample (normal approximation) results are given there. The results available in this module use exact calculations based on the enumeration of all possible values in the binomial distribution.

Suppose you have two populations from which dichotomous (binary) responses will be recorded. Assume without loss of generality that higher proportions are better. The probability (or risk) of cure in group 1 (the treatment group) is  $p_1$  and in group 2 (the reference group) is  $p_2$ . Random samples of  $n_1$  and  $n_2$  individuals are obtained from these two groups. The data from these samples can be displayed in a 2-by-2 contingency table as follows

Group	Success	Failure	Total
Treatment	$a$	$c$	$m$
Control	$b$	$d$	$n$
Totals	$s$	$f$	$N$

The following alternative notation is also used.

Group	Success	Failure	Total
Treatment	$x_{11}$	$x_{12}$	$n_1$
Control	$x_{21}$	$x_{22}$	$n_2$
Totals	$m_1$	$m_2$	$N$

The binomial proportions  $p_1$  and  $p_2$  are estimated from these data using the formulae

$$\hat{p}_1 = \frac{a}{m} = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{b}{n} = \frac{x_{21}}{n_2}$$



Let  $p_{1.0}$  represent the group 1 proportion tested by the null hypothesis  $H_0$ . The power of a test is computed at a specific value of the proportion which we will call  $p_{1.1}$ . Let  $\delta$  represent the smallest difference (margin of equivalence) between the two proportions that still results in the conclusion that the new treatment is equivalent to the current treatment. The set of statistical hypotheses that are tested is

$$H_0: |p_{1.0} - p_2| \geq \delta \text{ versus } H_1: |p_{1.0} - p_2| < \delta$$

These hypotheses can be rearranged to give

$$H_0: p_{1.0} - p_2 \leq \delta_L \text{ or } p_{1.0} - p_2 \geq \delta_U \text{ versus } H_1: \delta_L \leq p_{1.0} - p_2 \leq \delta_U$$

This composite hypothesis can be reduced to two one-sided hypotheses as follows

$$H_{0L}: p_{1.0} - p_2 \leq \delta_L \text{ versus } H_{1L}: \delta_L \leq p_{1.0} - p_2$$

$$H_{0U}: p_{1.0} - p_2 \geq \delta_U \text{ versus } H_{1U}: \delta_U \geq p_{1.0} - p_2$$

There are three common methods of specifying the margin of equivalence. The most direct is to simply give values for  $p_2$  and  $p_{1.0}$ . However, it is often more meaningful to give  $p_2$  and then specify  $p_{1.0}$  implicitly by reporting the difference, ratio, or odds ratio. Mathematically, the definitions of these parameterizations are

<u>Parameter</u>	<u>Computation</u>	<u>Alternative Hypotheses</u>
Difference	$\delta = p_{1.0} - p_2$	$H_1: \delta_L \leq p_{1.0} - p_2 \leq \delta_U$
Ratio	$\phi = p_{1.0} / p_2$	$H_1: \phi_L \leq p_{1.0} / p_2 \leq \phi_U$
Odds Ratio	$\psi = Odds_{1.0} / Odds_2$	$H_1: \psi_L \leq o_{1.0} / o_2 \leq \psi_U$

## Difference

The difference is perhaps the most direct method of comparison between two proportions. It is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One difficulty arises when the event of interest is rare. If a difference of 0.001 occurs when the baseline probability is 0.40, it would be dismissed as being trivial. However, if the baseline probability of a disease is 0.002, a 0.001 decrease would represent a reduction of 50%. Thus interpretation of the difference depends on the baseline probability of the event.

Note that  $\delta_L < 0$  and  $\delta_U > 0$ . Usually,  $\delta_L = -\delta_U$ .

## Equivalence using a Difference

The following example might help you understand the concept of an *equivalence* test. Suppose 60% of patients respond to the current treatment method ( $p_2 = 0.60$ ). If the response rate of the new treatment is no less than five percentage points better or worse than the existing treatment, it will be considered to be equivalent. Substituting these figures into the statistical hypotheses gives

$$H_0: p_{1.0} - p_2 \leq -0.05 \text{ or } p_{1.0} - p_2 \geq 0.05 \text{ versus } H_1: -0.05 \leq p_{1.0} - p_2 \leq 0.05$$

Using the relationship

$$p_{1.0} = p_2 + \delta$$

gives

$$H_0: p_{1.0} \leq 0.55 \text{ or } p_{1.0} \geq 0.65 \text{ versus } H_1: 0.55 \leq p_{1.0} \leq 0.65$$

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is between 0.55 and 0.65.

## Ratio

The ratio,  $\phi = p_{1.0} / p_2$ , gives the relative change in the probability of the response. Testing equivalence uses the formulation

$$H_0: p_{1.0} / p_2 \leq \phi_L \text{ or } p_{1.0} / p_2 \geq \phi_U \text{ versus } H_1: \phi_L \leq p_{1.0} / p_2 \leq \phi_U$$

The only subtlety is that for equivalence tests  $\phi_L < 1$  and  $\phi_U > 1$ . Usually,  $\phi_L = 1 / \phi_U$ .

## Equivalence using a Ratio

The following example might help you understand the concept of *equivalence* as defined by the ratio. Suppose that 60% of patients ( $p_2 = 0.60$ ) respond to the current treatment method. If the response rate of a new treatment is within 10% of 0.60, it will be considered to be equivalent to the standard treatment. Substituting these figures into the statistical hypotheses gives

$$H_0: p_{1.0} / p_2 \leq 0.9 \text{ or } p_{1.0} / p_2 \geq 1.1 \text{ versus } H_1: 0.9 \leq p_{1.0} / p_2 \leq 1.1$$

Using the relationship

$$p_{1.0} = \phi_0 p_2$$

gives

$$H_0: p_{1.0} \leq 0.54 \text{ or } p_{1.0} \geq 0.66 \text{ versus } H_1: 0.54 \leq p_{1.0} \leq 0.66$$

## Odds Ratio

The odds ratio,  $\psi = (p_{1.0} / (1 - p_{1.0})) / (p_2 / (1 - p_2))$ , gives the relative change in the odds ( $o$ ) of the response. Testing equivalence use the formulation

$$H_0: o_{1.0} / o_2 \leq \psi_L \text{ or } o_{1.0} / o_2 \geq \psi_U \text{ versus } H_1: \psi_L \leq o_{1.0} / o_2 \leq \psi_U$$

The only subtlety is that for equivalence tests  $\psi_L < 1$  and  $\psi_U > 1$ . Usually,  $\psi_L = 1 / \psi_U$ .

## Power Calculation

The power for a test statistic that is based on the normal approximation can be computed exactly using two binomial distributions. The following steps are taken to compute the power of these tests.

1. Find the critical values using the standard normal distribution. The critical values  $z_L$  and  $z_U$  are chosen as that value of  $z$  that leaves exactly the target value of alpha in the appropriate tail of the normal distribution.
2. Compute the value of the test statistic  $z_t$  for every combination of  $x_{11}$  and  $x_{21}$ . Note that  $x_{11}$  ranges from 0 to  $n_1$ , and  $x_{21}$  ranges from 0 to  $n_2$ . A small value (around 0.0001) can be added to the zero-cell counts to avoid numerical problems that occur when the cell value is zero.
3. If  $z_t > z_L$  and  $z_t < z_U$ , the combination is in the rejection region. Call all combinations of  $x_{11}$  and  $x_{21}$  that lead to a rejection the set  $A$ .
4. Compute the power for given values of  $p_{1.1}$  and  $p_2$  as

$$1 - \beta = \sum_A \binom{n_1}{x_{11}} p_{1.1}^{x_{11}} q_{1.1}^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_2^{x_{21}} q_2^{n_2 - x_{21}}$$

5. Compute the actual value of alpha achieved by the design by substituting  $p_{1.0L}$  and  $p_{1.0U}$  for  $p_{1.1}$  to obtain

$$\alpha_L = \sum_A \binom{n_1}{x_{11}} p_{1.0L}^{x_{11}} q_{1.0L}^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_2^{x_{21}} q_2^{n_2 - x_{21}}$$

and

$$\alpha_U = \sum_A \binom{n_1}{x_{11}} p_{1.0U}^{x_{11}} q_{1.0U}^{n_1 - x_{11}} \binom{n_2}{x_{21}} p_2^{x_{21}} q_2^{n_2 - x_{21}}$$

The value of alpha is then computed as the maximum of  $\alpha_L$  and  $\alpha_U$ .

## Asymptotic Approximations

When the values of  $n_1$  and  $n_2$  are large (say over 200), these formulas take a long time to evaluate. In this case, a large sample approximation can be used. The large sample approximation is made by replacing the values of  $\hat{p}_1$  and  $\hat{p}_2$  in the  $z$  statistic with the corresponding values of  $p_{1.1}$  and  $p_2$  and then computing the results based on the normal distribution. Note that in large samples, the Farrington and Manning statistic is substituted for the Gart and Nam statistic. Also, for large samples, the results for the odds ratio have not (to our knowledge) been published. In this case, we substitute the calculations which are based on the ratio hypotheses.

## Test Statistics

Several test statistics have been proposed for testing whether the difference, ratio, or odds ratio are different from a specified value. The main difference among the several test statistics is in the formula used to compute the standard error used in the denominator. These tests are based on the following  $z$ -test

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - c}{\hat{\sigma}}$$

The constant,  $c$ , represents a continuity correction that is applied in some cases. When the continuity correction is not used,  $c$  is zero. In power calculations, the values of  $\hat{p}_1$  and  $\hat{p}_2$  are not known. The corresponding values of  $p_{1,1}$  and  $p_2$  can be reasonable substitutes.

Following is a list of the test statistics available in **PASS**. The availability of several test statistics begs the question of which test statistic one should use. The answer is simple: one should use the test statistic that will be used to analyze the data. You may choose a method because it is a standard in your industry, because it seems to have better statistical properties, or because your statistical package calculates it. Whatever your reasons for selecting a certain test statistic, you should use the same test statistic when doing the analysis after the data have been collected.

### Z Test (Pooled)

This test was first proposed by Karl Pearson in 1900. Although this test is usually expressed directly as a chi-square statistic, it is expressed here as a  $z$  statistic so that it can be more easily used for one-sided hypothesis testing. The proportions are pooled (averaged) in computing the standard error. The formula for the test statistic is

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_1}$$

where

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

## Z Test (Unpooled)

This test statistic does not pool the two proportions in computing the standard error.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_2}$$

where

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

## Z Test with Continuity Correction (Pooled)

This test is the same as Z Test (Pooled), except that a continuity correction is used. Remember that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 + \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_1}$$

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\bar{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

where  $F$  is -1 for lower-tailed hypotheses and 1 for upper-tailed hypotheses.

## Z Test with Continuity Correction (Unpooled)

This test is the same as the Z Test (Unpooled), except that a continuity correction is used. Remember that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - \frac{F}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}{\hat{\sigma}_2}$$

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

where  $F$  is -1 for lower-tailed hypotheses and 1 for upper-tailed hypotheses.

## T-Test of Difference

Because of a detailed, comparative study of the behavior of several tests, D'Agostino (1988) and Upton (1982) proposed using the usual two-sample  $t$ -test for testing whether the two proportions are equal. One substitutes a '1' for a success and a '0' for a failure in the usual, two-sample  $t$ -test formula.

## Miettinen and Nurminen's Likelihood Score Test of the Difference

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the difference is equal to a specified, non-zero, value,  $\delta_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing this test statistic is

$$z_{MND} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_{MND}}$$

where

$$\hat{\sigma}_{MND} = \sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right) \left( \frac{N}{N-1} \right)}$$

$$\tilde{p}_1 = \tilde{p}_2 + \delta_0$$

$$\tilde{p}_1 = 2B \cos(A) - \frac{L_2}{3L_3}$$

$$A = \frac{1}{3} \left[ \pi + \cos^{-1} \left( \frac{C}{B^3} \right) \right]$$

$$B = \text{sign}(C) \sqrt{\frac{L_2^2}{9L_3} - \frac{L_1}{3L_3}}$$

$$C = \frac{L_2^3}{27L_3^3} - \frac{L_1 L_2}{6L_3^2} + \frac{L_0}{2L_3}$$

$$L_0 = x_{21} \delta_0 (1 - \delta_0)$$

$$L_1 = [N_2 \delta_0 - N - 2x_{21}] \delta_0 + M_1$$

$$L_2 = (N + N_2) \delta_0 - N - M_1$$

$$L_3 = N$$

### Miettinen and Nurminen's Likelihood Score Test of the Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the ratio is equal to a specified value,  $\phi_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 / \tilde{p}_2 = \phi_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{MNR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \phi_0^2 \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right) \left( \frac{N}{N-1} \right)}}$$

where

$$\begin{aligned}\tilde{p}_1 &= \tilde{p}_2 \phi_0 \\ \tilde{p}_2 &= \frac{-B - \sqrt{B^2 - 4AC}}{2A} \\ A &= N\phi_0 \\ B &= -[N_1\phi_0 + x_{11} + N_2 + x_{21}\phi_0] \\ C &= M_1\end{aligned}$$

### Miettinen and Nurminen's Likelihood Score Test of the Odds Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the odds ratio is equal to a specified value,  $\psi_0$ . Because the approach they used with the difference and ratio does not easily extend to the odds ratio, they used a score statistic approach for the odds ratio. The regular MLE's are  $\hat{p}_1$  and  $\hat{p}_2$ . The constrained MLE's are  $\tilde{p}_1$  and  $\tilde{p}_2$ . These estimates are constrained so that  $\tilde{\psi} = \psi_0$ . A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic. The formula for computing the test statistic is

$$z_{MNO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{N_2 \tilde{p}_1 \tilde{q}_1} + \frac{1}{N_2 \tilde{p}_2 \tilde{q}_2} \right) \left( \frac{N}{N-1} \right)}}$$

where

$$\begin{aligned}\tilde{p}_1 &= \frac{\tilde{p}_2 \psi_0}{1 + \tilde{p}_2(\psi_0 - 1)} \\ \tilde{p}_2 &= \frac{-B + \sqrt{B^2 - 4AC}}{2A} \\ A &= N_2(\psi_0 - 1) \\ B &= N_1\psi_0 + N_2 - M_1(\psi_0 - 1) \\ C &= -M_1\end{aligned}$$



## Farrington and Manning's Likelihood Score Test of the Difference

Farrington and Manning (1990) proposed a test statistic for testing whether the difference is equal to a specified value,  $\delta_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ , are used in the denominator. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMD} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

## Farrington and Manning's Likelihood Score Test of the Ratio

Farrington and Manning (1990) proposed a test statistic for testing whether the ratio is equal to a specified value,  $\phi_0$ . The regular MLE's,  $\hat{p}_1$  and  $\hat{p}_2$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_1$  and  $\tilde{p}_2$ , constrained so that  $\tilde{p}_1 / \tilde{p}_2 = \phi_0$ , are used in the denominator. A correction factor of  $N/(N-1)$  is applied to increase the variance estimate. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left( \frac{\tilde{p}_1 \tilde{q}_1}{n_1} + \phi_0^2 \frac{\tilde{p}_2 \tilde{q}_2}{n_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

## Farrington and Manning's Likelihood Score Test of the Odds Ratio

Farrington and Manning (1990) indicate that the Miettinen and Nurminen statistic may be modified by removing the factor  $N/(N-1)$ .

The formula for computing this test statistic is

$$z_{FMO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{N_2 \tilde{p}_1 \tilde{q}_1} + \frac{1}{N_2 \tilde{p}_2 \tilde{q}_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

### Gart and Nam's Likelihood Score Test of the Difference

Gart and Nam (1990), page 638, proposed a modification to the Farrington and Manning (1988) difference test that corrects for skewness. Let  $z_{FMD}(\delta)$  stand for the Farrington and Manning difference test statistic described above. The skewness-corrected test statistic,  $z_{GND}$ , is the appropriate solution to the quadratic equation

$$(-\tilde{\gamma})z_{GND}^2 + (-1)z_{GND} + (z_{FMD}(\delta) + \tilde{\gamma}) = 0$$

where

$$\tilde{\gamma} = \frac{\tilde{V}^{3/2}(\delta)}{6} \left( \frac{\tilde{p}_1 \tilde{q}_1 (\tilde{q}_1 - \tilde{p}_1)}{n_1^2} - \frac{\tilde{p}_2 \tilde{q}_2 (\tilde{q}_2 - \tilde{p}_2)}{n_2^2} \right)$$

### Gart and Nam's Likelihood Score Test of the Ratio

Gart and Nam (1988), page 329, proposed a modification to the Farrington and Manning (1988) ratio test that corrects for skewness. Let  $z_{FMR}(\phi)$  stand for the Farrington and Manning ratio test statistic described above. The skewness-corrected test statistic,  $z_{GNR}$ , is the appropriate solution to the quadratic equation

$$(-\tilde{\varphi})z_{GNR}^2 + (-1)z_{GNR} + (z_{FMR}(\phi) + \tilde{\varphi}) = 0$$

where

$$\tilde{\varphi} = \frac{1}{6\tilde{u}^{3/2}} \left( \frac{\tilde{q}_1 (\tilde{q}_1 - \tilde{p}_1)}{n_1^2 \tilde{p}_1^2} - \frac{\tilde{q}_2 (\tilde{q}_2 - \tilde{p}_2)}{n_2^2 \tilde{p}_2^2} \right)$$

$$\tilde{u} = \frac{\tilde{q}_1}{n_1 \tilde{p}_1} + \frac{\tilde{q}_2}{n_2 \tilde{p}_2}$$

## 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 Procedure Templates.

## Data Tab (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers four procedures, each of which has different options. This section documents options that are common to all four procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *P1.1*, *Alpha*, *Beta*, *N1*, and *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.

### Test Statistic

Specify which test statistic is used in searching and reporting. Although the pooled *z*-test is commonly shown in elementary statistics books, the likelihood score test is arguably the best choice.

Note that *C.C.* is an abbreviation for *Continuity Correction*. This refers to the adding or subtracting  $1/(2n)$  to (or from) the numerator of the *z*-value to bring the normal approximation closer to the binomial distribution.

## P2 (Reference Group Proportion)

Specify the value of P2, the reference, baseline, or control group's proportion. The null hypothesis is that the two proportions differ by no more than a specified amount. Since P2 is a proportion, these values must be between 0 and 1.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

## N1 (Sample Size Group 1)

Enter a value (or range of values) for the sample size of this group. 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  $N2 = N1$ .

## 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 unequal proportions when in fact they are not equal.

Values must be between 0 and 1. Because this is a two-sided test, unless you have good reason to do otherwise, the value of 0.05 is recommended. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.025 0.05* or *0.01 to 0.10 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 unequal proportions when in fact they are equivalent.

Values must be between 0 and 1. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions,  $P_1$  and  $P_2$ .  $P_{1.0}$  is the value of the  $P_1$  assumed by the null hypothesis and  $P_{1.1}$  is the value of  $P_1$  at which the power is calculated.

### **$P_{1.0U}$ & $P_{1.0L}$ (Upper & Lower Equivalence Proportion)**

Specify the *margin of equivalence* directly by giving the upper and lower bounds of  $P_{1.0}$ . The two groups are assumed to be equivalent when  $P_{1.0}$  is between these values. Thus,  $P_{1.0U}$  should be greater than  $P_2$  and  $P_{1.0L}$  should be less than  $P_2$ .

Note that the values of  $P_{1.0U}$  and  $P_{1.0L}$  are used in pairs. Thus, the first values of  $P_{1.0U}$  and  $P_{1.0L}$  are used together, then the second values of each are used, and so on.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

Proportions must be between 0 and 1. They cannot take on the values 0 or 1. These values should surround  $P_2$ .

### **$P_{1.1}$ (Actual Proportion)**

This option specifies the value of  $P_{1.1}$ , which is the value of the treatment proportion at which the power is to be calculated. Proportions must be between 0 and 1. They cannot take on the values 0 or 1.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $D0$ , and  $D1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### D0.U & D0.L (Upper & Lower Equivalence Difference)

Specify the *margin of equivalence* by specifying the largest distance above ( $D0.U$ ) and below ( $D0.L$ )  $P2$  which will still result in the conclusion of equivalence. As long as the actual difference is between these two values, the difference is not considered to be large enough to be of practical importance.

The values of  $D0.U$  must be positive and the values of  $D0.L$  must be negative.  $D0.L$  can be set to  $^{-D0.U}$ , which is usually what is desired.

The power calculations assume that  $P1.0$  is the value of  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0U = D0.U + P2$ .

You may enter a range of values for  $D0.U$  such as  $.03 .05 .10$  or  $.05 to .20 by .05$ .

Note that if you enter values for  $D0.L$  (other than  $^{-D0.U}$ ), they are used in pairs with the values of  $D0.U$ . Thus, the first values of  $D0.U$  and  $D0.L$  are used together, then the second values of each are used, and so on.

RANGE:

$D0.L$  must be between  $-1$  and  $0$ .  $D0.U$  must be between  $0$  and  $1$ . Neither can take on the values  $-1$ ,  $0$ , or  $1$ .

### D1 (Actual Difference)

This option specifies the actual difference between  $P1.1$  (the actual value of  $P1$ ) and  $P2$ . This is the value of the difference at which the power is calculated. In equivalence trials, this difference is often set to  $0$ .

The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group). This difference is used with  $P2$  to calculate the true value of  $P1$  using the formula:  $P1.1 = D1 + P2$ .

You may enter a range of values such as  $-.05 0 .5$  or  $-.05 to .05 by .02$ . Actual differences must be between  $-1$  and  $1$ . They cannot take on the values  $-1$  or  $1$ .

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $R0$ , and  $R1$  are given, the values of  $P1.0$  and  $P1.1$  can be calculated.

### R0.U & R0.L (Upper & Lower Equivalence Ratio)

Specify the *margin of equivalence* by specifying the largest ratio ( $P1/P2$ ) above ( $R0.U$ ) and below ( $R0.L$ ) 1 which will still result in the conclusion of equivalence. As long as the actual ratio is between these two values, the difference between the proportions is not considered to be large enough to be of practical importance.

The values of  $R0.U$  must be greater than 1 and the values of  $R0.L$  must be less than 1.  $R0.L$  can be set to ' $1/R0.U$ ,' which is often desired.

The power calculations assume that  $P1.0$  is the value of  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0U = R0.U \times P2$ .

You may enter a range of values for  $R0.U$  such as *1.1 1.5 1.8* or *1.1 to 2.1 by 0.2*.

Note that if you enter values for  $R0.L$  (other than ' $1/R0.U$ '), they are used in pairs with the values of  $R0.U$ . Thus, the first values of  $R0.U$  and  $R0.L$  are used together, then the second values of each are used, and so on.

$R0.L$  must be between 0 and 1.  $R0.U$  must be greater than 1. Neither can take on the value 1.

### R1 (Actual Ratio)

This option specifies the ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the actual proportion in the treatment group. The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1. This difference is used with  $P2$  to calculate the value of  $P1$  using the formula:  $P1.1 = R1 \times P2$ . In equivalence trials, this ratio is often set to 1.

Ratios must be positive. You may enter a range of values such as *0.95 1 1.05* or *0.9 to 1.9 by 0.02*.

## Data Tab (Odds Ratio)

This section documents options that are used when the parameterization is in terms of the odds ratios,  $O1.1 / O2$  and  $O1.0 / O2$ . Note that the odds are defined as  $O2 = P2 / (1 - P2)$ ,  $O1.0 = P1.0 / (1 - P1.0)$ , etc.  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $OR0$ , and  $OR1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

## OR0.U & OR0.L (Upper & Lower Equivalence Odds Ratio)

Specify the *margin of equivalence* by specifying the largest odds ratio above ( $OR0.U$ ) and below ( $OR0.L$ ) 1 which will still result in the conclusion of equivalence. As long as the actual odds ratio is between these two values, the difference between the proportions is not large enough to be of practical importance.

The values of  $OR0.U$  must be greater than 1 and the values of  $OR0.L$  must be less than 1.  $OR0.L$  can be set to ' $1/OR0.U$ ,' which is often desired.

The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$ .

You may enter a range of values for  $OR0.U$  such as *1.1 1.5 1.8* or *1.1 to 2.1 by 0.2*.

Note that if you enter values for  $OR0.L$  (other than ' $1/OR0.U$ '), they are used in pairs with the values of  $OR0.U$ . Thus, the first values of  $OR0.U$  and  $OR0.L$  are used together, next the second values of each are used, and so on.

$OR0.L$  must be between 0 and 1.  $OR0.U$  must be greater than 1. Neither can take on the value 1.

## OR1 (Actual Odds Ratio)

This option specifies the odds ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the actual proportion in the treatment group. The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1. This value is used with  $P2$  to calculate the value of  $P1$ . In equivalence trials, this odds ratio is often set to 1.

Odds ratios must be positive. You may enter a range of values such as *0.95 1 1.05* or *0.9 to 1.9 by 0.02*.



## Options Tab

The Options tab contains various limits and options.

### Maximum N1 or N2 Exact

When either N1 or N2 is above this amount, power calculations are based on the normal approximation to the binomial. When the normal approximation to the binomial is used, the actual value of alpha is not calculated. Currently, for three-gigahertz computers, a value near 200 is reasonable. As computers increase in speed, this number may be increased.

### Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

### Zero Count Adjustment Method

Zero cell counts cause many calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option 'Add to zero cells only.'

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts. We have found that adding 0.0001 seems to work well.

### Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This value indicates the amount that is added. We have found that 0.0001 works well.

Be warned that the value of the ratio and the odds ratio will be affected by the amount specified here!

## Example1 - Finding Power

A study is being designed to establish the equivalence of a new treatment compared to the current treatment. Historically, the current treatment has enjoyed a 50% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the current treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the current treatment. The researchers will recommend adoption of the new treatment if its cure rate is within 15% of the standard treatment.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data. They want to study the power of the Farrington and Manning test at group sample sizes ranging from 50 to 500 for detecting a difference inside 15% when the actual cure rate of the new treatment ranges from 50% to 60%. The significance level will be 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.15</b>
D0.L .....	<b>-D0.U</b>
D1 .....	<b>0.00 0.05 0.10</b>
P2.....	<b>0.5</b>
N1 .....	<b>50 to 500 by 50</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100</b>

### Annotated Output

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

## Numeric Results

Numeric Results for Equivalence Tests Based on the Difference: P1 - P2

H0: P1-P2≤D0.L or P1-P2≥D0.U. H1: D0.L<P1-P2=D1<D0.U.

Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Target Alpha	Actual Alpha
Power	N1	N2								
0.0000	50	50	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	0.0515
0.3793	100	100	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	0.0489
0.6689	150	150	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
0.8305	200	200	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
0.9160	250	250	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
0.9594	300	300	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
0.9808	350	350	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
0.9911	400	400	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
Report continues ...										

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 100.

### Report Definitions

'Power' is the probability of rejecting a false null hypothesis. It should be close to one.

'Power' is the probability of concluding equivalence when equivalence is correct.

'Beta' is the probability of accepting a false H0. Beta = 1 - Power.

'N1 and N2' are the sizes of the samples drawn from the corresponding groups.

'P2' is the response rate for group two which is the standard, reference, baseline, or control group.

'P1.0L' is the smallest treatment-group response rate that still yields an equivalence conclusion.

'P1.0U' is the largest treatment-group response rate that still yields an equivalence conclusion.

'D0.L' is the lowest difference that still results in the conclusion of equivalence.

'D0.U' is the highest difference that still results in the conclusion of equivalence.

'D1' is the actual difference, P1-P2, at which the power is calculated.

'Target Alpha' is the probability of rejecting a true null hypothesis that was desired.

'Actual Alpha' is the value of alpha that is actually achieved. Only available for exact results.

'Grp 1' refers to Group 1 which is the treatment or experimental group.

'Grp 2' refers to Group 2 which is the reference, standard, or control group.

'Equiv.' refers to a small amount that is not of practical importance.

'Actual' refers to the true value at which the power is computed.

### Summary Statements

Sample sizes of 50 in the treatment group and 50 in the reference group achieve 0% power to detect equivalence. The margin of equivalence, given in terms of the difference, extends from -0.1500 to 0.1500. The actual difference is 0.0000. The reference group proportion is 0.5000. The calculations assume that two, one-sided likelihood score (Farrington & Manning) tests are used. Although the significance level is targeted at 0.0500, the level actually achieved is 0.0515.

This report shows the values of each of the parameters, one scenario per row. Note that the actual alpha value is blank for sample sizes greater than 100, which was the limit set for exact computation.

Most of the report columns have obvious interpretations. Those that may not be obvious are presented here.

## Prop Grp 2 P2

This is the value of P2, the response rate in the control group.

## Lower & Upper Equiv. Grp 1 Prop: P1.0L & P1.0U

These are the margin of equivalence for the response rate of the treatment group as specified by the null hypothesis of non-equivalence. Values of P1 inside these limits are considered equivalent to P2.

## Lower & Upper Equiv. Margin Diff: D0.L & D0.U

These set the margin of equivalence for the difference in response rates. Values of the difference outside these limits are considered *non-equivalent*.

## Actual Margin Diff D1

This is the value of D1, the difference between the two group proportions at which the power is computed. This is the value of the difference under the alternative hypothesis.

## Target Alpha

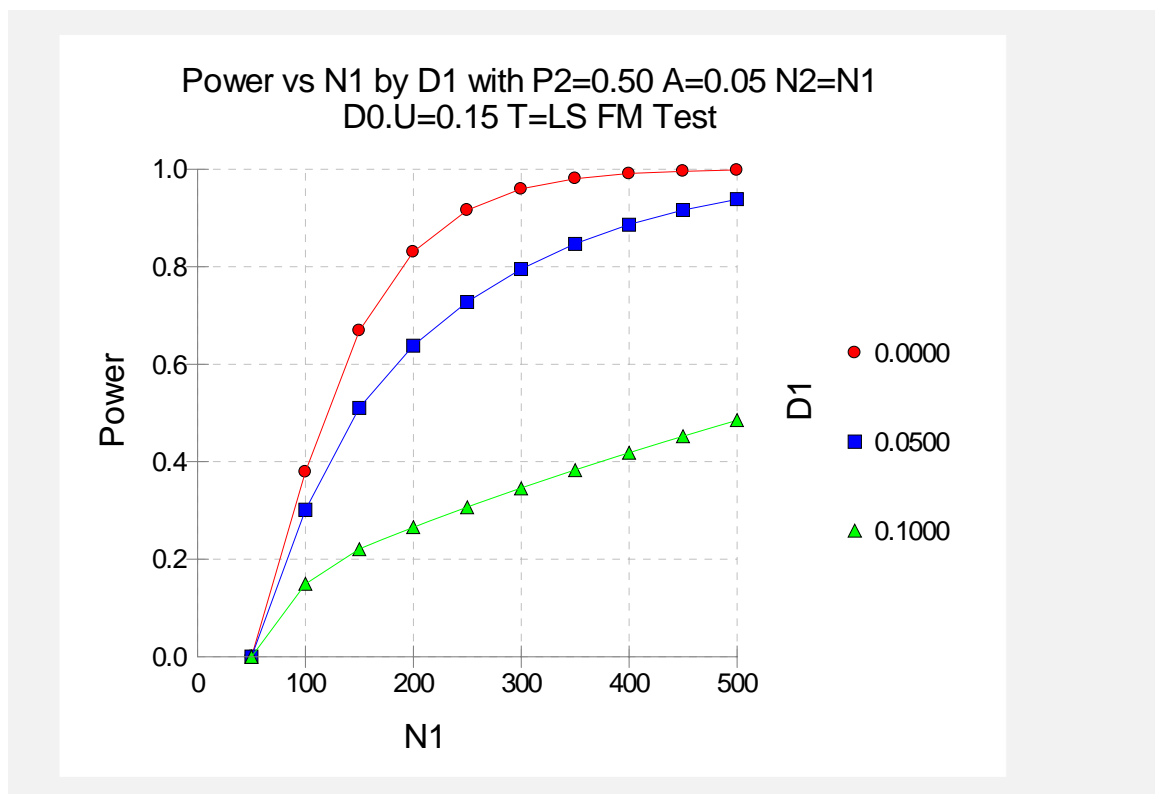
This is the value of alpha that was targeted by the design. Note that the target alpha is not usually achieved exactly. For two-sided tests, this value will usually be 0.05.

## Actual Alpha

This is the value of alpha that was actually achieved by this design. Note that since the limit on exact calculations was set to 100, and since this value is calculated exactly, it is not shown for values of N1 greater than 100.

The difference between the Target Alpha and the Actual Alpha is caused by the discrete nature of the binomial distribution and the use of the normal approximation to the binomial in determining the critical value of the test statistic.

## Plots Section



The values from the table are displayed in the above chart. This chart gives us a quick look at the sample size that will be required for various values of D1.

## Example 2 - Finding the Sample Size

Continuing with the scenario given in Example 1, the researchers want to determine the sample size necessary for each value of D1 to achieve a power of 0.80. To cut down on the runtime, they decide to look at approximate values whenever N1 is greater than 100.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.15</b>
D0.L.....	<b>-D0.U</b>
D1.....	<b>0.00 0.05 0.10</b>
P2.....	<b>0.5</b>
N1.....	<i>Ignored since this is the Find setting.</i>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100</b>

### Annotated Output

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

### Numeric Results

Numeric Results for Equivalence Tests Based on the Difference: P1 - P2  
H0:  $P1 - P2 \leq D0.L$  or  $P1 - P2 \geq D0.U$ . H1:  $D0.L < P1 - P2 = D1 < D0.U$ .  
Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Target Alpha	Actual Alpha
<b>Power</b>	<b>N1</b>	<b>N2</b>								
0.8003	188	188	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0000	0.0500	
0.8001	304	304	0.5000	0.3500	0.6500	-0.1500	0.1500	0.0500	0.0500	
0.8001	1202	1202	0.5000	0.3500	0.6500	-0.1500	0.1500	0.1000	0.0500	

The required sample size will depend a great deal on the value of D1. Any effort spent determining an accurate value for D1 will be worthwhile.

## Example3 – Comparing the Power of Several Test Statistics

Continuing with Example 1, the researchers want to determine which of the eight possible test statistics to adopt by using the comparative reports and charts that *PASS* produces. They decide to compare the powers and actual alphas for various sample sizes between 50 and 200 when D1 is 0.1.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.15</b>
D0.L .....	<b>-D0.U</b>
D1 .....	<b>0.10</b>
P2 .....	<b>0.5</b>
N1 .....	<b>50 to 200 by 50</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>300</b>
<b>Reports Tab</b>	
Show Numeric Report .....	<b>Not checked</b>
Show Comparative Reports .....	<b>Checked</b>
Show Definitions .....	<b>Not checked</b>
Show Plots .....	<b>Not checked</b>
Show Comparative Plots .....	<b>Checked</b>
Summary Statement Rows .....	<b>0</b>

### Annotated Output

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

## Numeric Results

### Power Comparison of Equivalence Tests Based on the Difference: $P_1 - P_2$

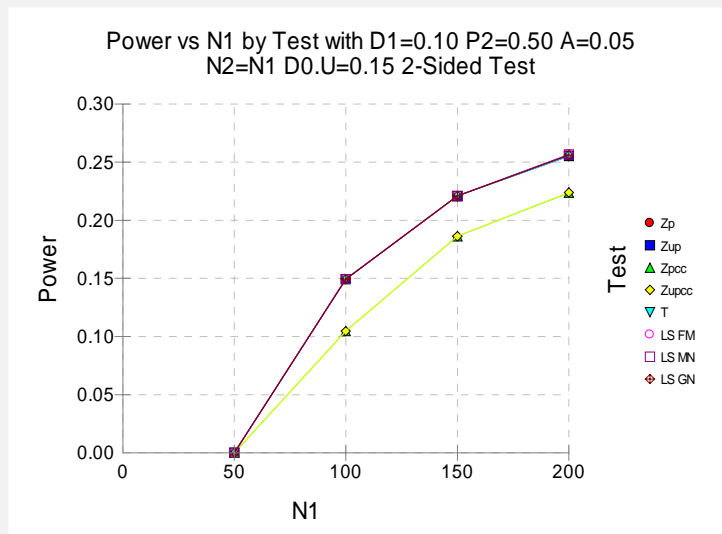
$H_0: P_1 - P_2 \leq D_{0.L}$  or  $P_1 - P_2 \geq D_{0.U}$ .  $H_1: D_{0.L} < P_1 - P_2 = D_1 < D_{0.U}$ .

N1/N2	P2	Upper Equiv. Margin D0.U	Target Alpha	Z(P) Test Power	Z(UnP) Test Power	Z(P) CC Test Power	Z(UnP) CC Test Power	T Test Power	F.M. Score Power	M.N. Score Power	G.N. Score Power
50/50	0.5000	0.1500	0.0500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100/100	0.5000	0.1500	0.0500	0.1494	0.1494	0.1047	0.1047	0.1493	0.1495	0.1494	0.1494
150/150	0.5000	0.1500	0.0500	0.2208	0.2208	0.1863	0.1863	0.2208	0.2208	0.2208	0.2208
200/200	0.5000	0.1500	0.0500	0.2552	0.2553	0.2238	0.2239	0.2552	0.2566	0.2566	0.2560

### Actual Alpha Comparison of Equivalence Tests Based on the Difference: $P_1 - P_2$

$H_0: P_1 - P_2 \leq D_{0.L}$  or  $P_1 - P_2 \geq D_{0.U}$ .  $H_1: D_{0.L} < P_1 - P_2 = D_1 < D_{0.U}$ .

N1/N2	P2	Upper Equiv. Margin D0.U	Target Alpha	Z(P) Test Alpha	Z(UnP) Test Alpha	Z(P) CC Test Alpha	Z(UnP) CC Test Alpha	T Test Alpha	F.M. Score Alpha	M.N. Score Alpha	G.N. Score Alpha
50/50	0.5000	0.1500	0.0500	0.0515	0.0515	0.0334	0.0334	0.0514	0.0515	0.0515	0.0515
100/100	0.5000	0.1500	0.0500	0.0486	0.0486	0.0358	0.0358	0.0485	0.0489	0.0487	0.0487
150/150	0.5000	0.1500	0.0500	0.0495	0.0495	0.0386	0.0386	0.0495	0.0495	0.0495	0.0495
200/200	0.5000	0.1500	0.0500	0.0465	0.0468	0.0376	0.0378	0.0465	0.0488	0.0488	0.0481



It is interesting to note that the powers of the continuity-corrected test statistics are consistently lower than the other tests. This occurs because the actual alpha achieved by these tests is lower than for the other tests. An interesting finding of this example is that the regular  $t$ -test performed about as well as the  $z$ -test.

## Example4 - Validation using Chow with Equal Sample Sizes

Chow et al. (2003), page 91, present a sample size study in which  $P2 = 0.75$ ,  $D0.U = 0.2$ ,  $D0.L = -0.2$ ,  $D1 = 0.05$ ,  $\alpha = 0.05$ , and  $\beta = 0.2$ . Using the pooled Z test statistic, they found the sample size to be 96 in each group.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic .....	<b>Z Test (Pooled)</b>
Specify Treatment Proportion using..	<b>Differences (P1-P2)</b>
D0.U .....	<b>0.2</b>
D0.L .....	<b>-D0.U</b>
D1 .....	<b>0.05</b>
P2 .....	<b>0.75</b>
N1 .....	<i>Ignored since this is the Find setting.</i>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.20</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2 (Set low for a rapid search.)</b>

### Annotated Output

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



## Numeric Results

Numeric Results for Equivalence Tests Based on the Difference:  $P1 - P2$

$H_0: P1 - P2 \leq D0.L$  or  $P1 - P2 \geq D0.U$ .  $H_1: D0.L < P1 - P2 < D0.U$ .

Test Statistic: Z test (pooled)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Target Alpha	Actual Alpha
Power	N1	N2								
0.8028	98	98	0.7500	0.5500	0.9500	-0.2000	0.2000	0.0500	0.0500	

*PASS* found the required sample size to be 98 which is slightly larger than the 96 that Chow obtained. This is mainly due to the rounding to two decimal places that Chow did in this example. We used the exact option in *PASS* and obtained  $N1 = 99$ . Thus, *PASS* was indeed closer than was Chow.

## Example5 - Validation using Tuber-Bitter with Equal Sample Sizes

Tuber-Bitter et al. (2000), page 1271, present a sample size study in which  $P_2 = 0.1$ ;  $D0.U = 0.01, 0.02, 0.03$ ;  $D0.L = -D0.U$ ;  $D1 = 0.0$ ;  $\alpha = 0.05$ ; and  $\beta = 0.1$ . Using the pooled Z test statistic, they found the sample sizes to be 19484, 4871, and 2165 in each group.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic .....	<b>Z Test (Pooled)</b>
Specify Treatment Proportion using..	<b>Differences (P1-P2)</b>
D0.U .....	<b>.01 .02 .03</b>
D0.L .....	<b>-D0.U</b>
D1 .....	<b>0.0</b>
P2 .....	<b>0.1</b>
N1 .....	<i>Ignored since this is the Find setting.</i>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>2 (Set low for a rapid search.)</b>

### Annotated Output

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

### Numeric Results

Numeric Results for Equivalence Tests Based on the Difference:  $P_1 - P_2$   
 $H_0: P_1 - P_2 \leq D0.L \text{ or } P_1 - P_2 \geq D0.U$ .  $H_1: D0.L < P_1 - P_2 < D1 < D0.U$ .  
 Test Statistic: Z test (pooled)

	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 P2	Lower Trivial Grp 1 Prop P1.0L	Upper Trivial Grp 1 Prop P1.0U	Lower Trivial Margin Diff D0.L	Upper Trivial Margin Diff D0.U	Actual Margin Diff D1	Target Alpha	Actual Alpha
Power										
0.9000	19480	19480	0.1000	0.0900	0.1100	-0.0100	0.0100	0.0000	0.0500	
0.9000	4870	4870	0.1000	0.0800	0.1200	-0.0200	0.0200	0.0000	0.0500	
0.9001	2165	2165	0.1000	0.0700	0.1300	-0.0300	0.0300	0.0000	0.0500	

**PASS** found the required sample sizes to within rounding error of Tuber-Bitter.

## Example6 - Computing the Power after Completing an Experiment

Researchers are testing a generic drug to determine if it is equivalent to the name-brand alternative. Equivalence is declared if the success rate of the generic brand is no more than 5% from that of the name-brand drug. In a study with 1000 individuals in each group, they find that 774, or 77.4%, are successfully treated using the name-brand drug, and 700, or 70%, respond to the generic drug. An equivalence test (exact test) with  $\alpha = 0.05$  failed to declare that the two drugs are equivalent. The researchers would now like to compute the power for actual differences ranging from 0 to 4%. Suppose that the true value for the response rate for the name-brand drug is 77%.

Note that the power is not calculated at the difference observed in the study, 77.4%. In fact, the difference observed in the study is larger than the proposed equivalence difference, 5%. It would make no sense to perform a power calculation for a difference larger than the equivalence difference. It is more informative to study a range of values smaller than or equal to the equivalence difference.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.05</b>
D0.L.....	<b>-D0.U</b>
D1 .....	<b>0.00 to 0.04 by 0.01</b>
P2.....	<b>0.77</b>
N1 .....	<b>1000</b>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
<b>Options Tab</b>	
Maximum N1 or N2 Exact .....	<b>100</b>

### Annotated Output

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

## Numeric Results

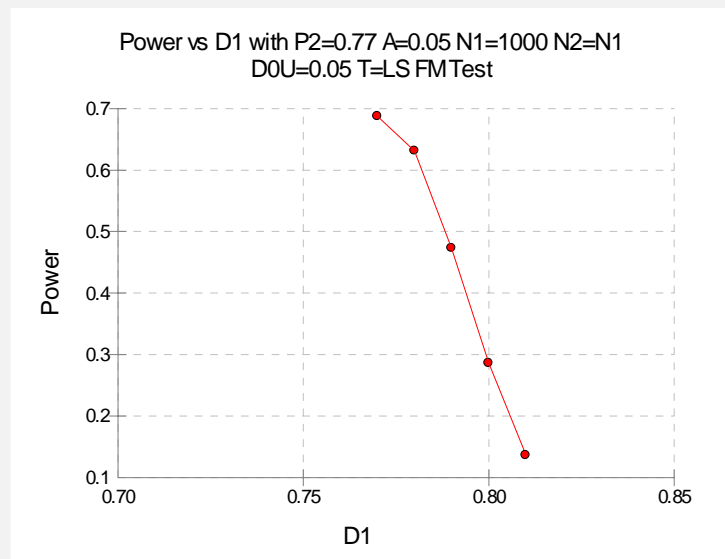
Numeric Results for Equivalence Tests Based on the Difference:  $P_1 - P_2$

$H_0: P_1 - P_2 \leq D_{0.L} \text{ or } P_1 - P_2 \geq D_{0.U}$ .  $H_1: D_{0.L} < P_1 - P_2 < D_{0.U}$ .

Test Statistic: Score test (Farrington & Manning)

	Sample Size Grp 1	Sample Size Grp 2	Prop Grp 2 P2	Lower Trivial Grp 1 Prop P1.0L	Upper Trivial Grp 1 Prop P1.0U	Lower Trivial Margin Diff D0.L	Upper Trivial Margin Diff D0.U	Actual Margin Diff D1	Target Alpha	Actual Alpha
Power	N1	N2								
0.6875	1000	1000	0.7700	0.7200	0.8200	-0.0500	0.0500	0.0000	0.0500	
0.6313	1000	1000	0.7700	0.7200	0.8200	-0.0500	0.0500	0.0100	0.0500	
0.4731	1000	1000	0.7700	0.7200	0.8200	-0.0500	0.0500	0.0200	0.0500	
0.2857	1000	1000	0.7700	0.7200	0.8200	-0.0500	0.0500	0.0300	0.0500	
0.1362	1000	1000	0.7700	0.7200	0.8200	-0.0500	0.0500	0.0400	0.0500	

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 100.



The power of the test ranges from 68.75% if the true difference is actually 0.0% to 13.62% if the true difference is 4%.

## Example7 - Finding the Sample Size using Proportions

A study is being designed to prove the equivalence of a new drug to the current standard. The current drug is effective in 85% of cases. The new drug, however, is cheaper to produce. The new drug will be deemed equivalent to the standard if its success rate is between 78% and 92%. What sample sizes are necessary to obtain 80% or 90% power for actual success rates ranging from 80% to 90%? The researchers will test at a significance level of 0.05 using the Farrington and Manning likelihood score test.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the proportion parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
P1.0U .....	<b>0.92</b>
P1.0L.....	<b>0.78</b>
P1.1 .....	<b>0.80 to 0.90 by 0.02</b>
P2.....	<b>0.85</b>
N1.....	<i>Ignored since this is the Find setting</i>
N2.....	<b>Use R</b>
R.....	<b>1.0</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20 0.10</b>
<b>Equivalence Tab</b>	

### Options Tab

Maximum N1 or N2 Exact ..... **100**

### Annotated Output

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

## Numeric Results

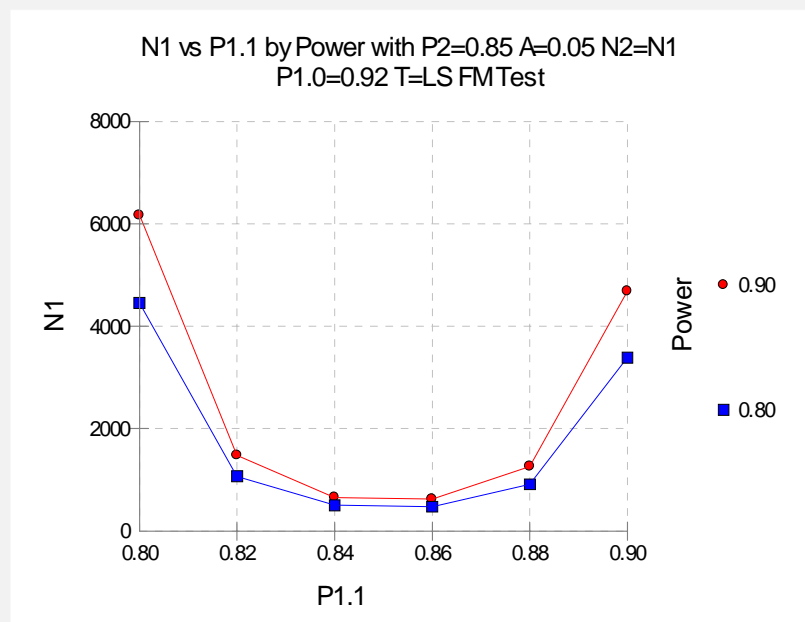
Numeric Results for Equivalence Tests Based on the Difference:  $P_1 - P_2$

$H_0: P_1 - P_2 \leq D_{0.L}$  or  $P_1 - P_2 \geq D_{0.U}$ .  $H_1: D_{0.L} < P_1 - P_2 < D_{0.U}$ .

Test Statistic: Score test (Farrington & Manning)

Power	Sample Size Grp 1 N1	Sample Size Grp 2 N2	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Target Alpha	Actual Alpha
0.9000	6166	6166	0.8500	0.7800	0.9200	-0.0700	0.0700	-0.0500	0.0500	
0.8001	4453	4453	0.8500	0.7800	0.9200	-0.0700	0.0700	-0.0500	0.0500	
0.9000	1480	1480	0.8500	0.7800	0.9200	-0.0700	0.0700	-0.0300	0.0500	
0.8002	1070	1070	0.8500	0.7800	0.9200	-0.0700	0.0700	-0.0300	0.0500	
0.9001	655	655	0.8500	0.7800	0.9200	-0.0700	0.0700	-0.0100	0.0500	
0.8008	503	503	0.8500	0.7800	0.9200	-0.0700	0.0700	-0.0100	0.0500	
0.9004	622	622	0.8500	0.7800	0.9200	-0.0700	0.0700	0.0100	0.0500	
0.8004	477	477	0.8500	0.7800	0.9200	-0.0700	0.0700	0.0100	0.0500	
0.9002	1261	1261	0.8500	0.7800	0.9200	-0.0700	0.0700	0.0300	0.0500	
0.8002	912	912	0.8500	0.7800	0.9200	-0.0700	0.0700	0.0300	0.0500	
0.9000	4685	4685	0.8500	0.7800	0.9200	-0.0700	0.0700	0.0500	0.0500	
0.8000	3386	3386	0.8500	0.7800	0.9200	-0.0700	0.0700	0.0500	0.0500	

Note: exact results based on the binomial were only calculated when both N1 and N2 were less than 100.



It is evident from these results that the sample sizes required to achieve 80% and 90% power depend a great deal on the actual value of the success rate,  $P_{1.1}$ .

## Chapter 220

# Group Sequential Tests of Two Proportions

## 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 group proportions. Other modules perform similar analyses for the comparison of means 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:

$$z_k = \frac{\hat{p}_{1k} - \hat{p}_{2k}}{\sqrt{\hat{p}_{1k}(1 - \hat{p}_{1k}) + \hat{p}_{2k}(1 - \hat{p}_{2k})}}$$

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 assumed to be normally distributed.

## Spending Functions

Lan and DeMets (1983) introduced alpha spending functions,  $\alpha(\tau)$ , that determine a set of boundaries  $b_1, b_2, \dots, b_K$  for the sequence of test statistics  $z_1, z_2, \dots, z_K$ . 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  $\tau$  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,  $\tau$  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  $\alpha$  level when the trial is complete. That is,

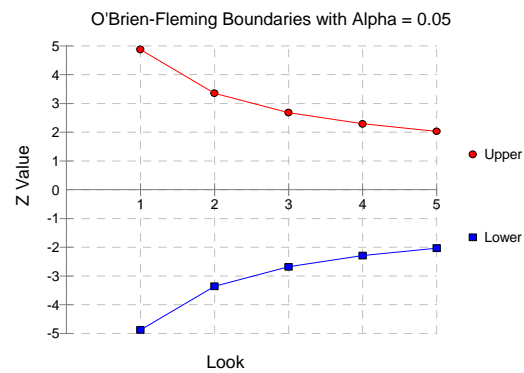
$$\Pr(|z_1| \geq b_1 \text{ or } |z_2| \geq b_2 \text{ or } \dots \text{ or } |z_K| \geq b_K) = \alpha(\tau)$$

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.

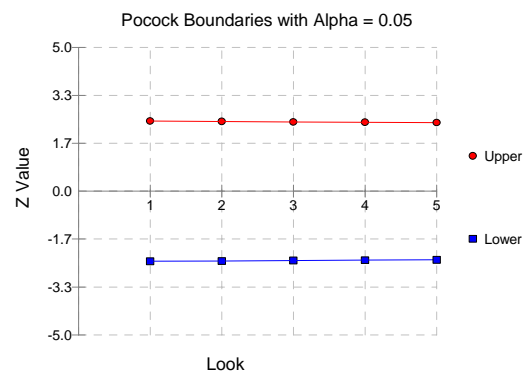
**PASS** provides five popular spending functions plus the ability to enter and analyze your own boundaries. These are calculated as follows:



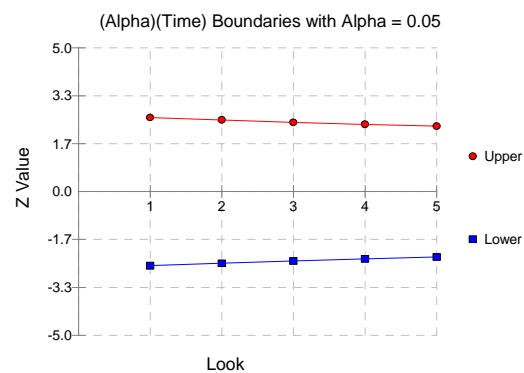
1. O'Brien-Fleming  $2 - 2\Phi\left(\frac{Z_{\alpha/2}}{\sqrt{\tau}}\right)$



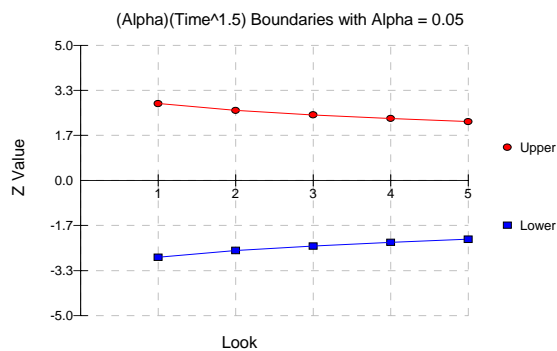
2. Pocock  $\alpha \ln(1 + (e - 1)\tau)$



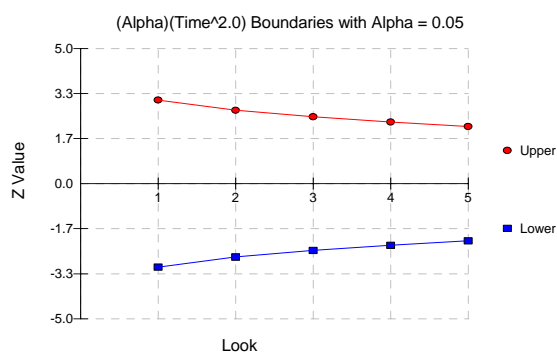
3. Alpha \* time  $\alpha\tau$



### 4. Alpha \* time<sup>1.5</sup> $\alpha\tau^{3/2}$



### 5. Alpha \* time<sup>2</sup> $\alpha\tau^2$



### 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 proportions 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  $\theta$ . The test statistics  $z_k$  follow the multivariate normal distribution with means  $\theta\sqrt{\tau_k}$  and, for  $j \leq k$ , covariances  $\sqrt{\tau_k / \tau_j}$ . The drift parameter is related to the parameters of the z-test through the equation

$$\theta = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{N1} + \frac{\bar{p}(1-\bar{p})}{N2}}}$$

where

$$\bar{p} = \frac{N1p_1 + N2p_2}{N1 + N2}$$

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 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 proportions, 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 *P1*, *P2*, *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.

## P1 (Proportion in Group 1)

Enter value(s) for the response proportion in the first group under both hypotheses and the response proportion of the second group under the null hypothesis of equal proportions. The values must be between zero and one.

You may enter a range of values such as *0.1*, *0.2*, *0.3* or *0.1 to 0.9 by 0.2*.

## P2 (Proportion in Group 2)

Enter value(s) for the response proportion of the second group under the alternative hypothesis. You may enter a range of values such as *0.1*, *0.2*, *0.3* or *0.1 to 0.9 by 0.2*.

## 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 proportions when in fact the proportions 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 proportions 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, respectively.

## 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 unless you can justify using a one-sided 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  $[Y]$  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  $N2 = N1$ .

## Use Continuity Correction

Specify whether to use the Continuity Correction. This option applies an adjustment to the sample sizes that is recommended by Fleiss(1981) page 45 to make the alpha and beta values more accurate. The formula for the adjustment is

$$N1_{new} = \frac{N1_{old}}{4} \left( 1 + \sqrt{1 + \frac{2(R+1)}{R(N1_{old})|P1-P2|}} \right)^2$$

## 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 were 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 0.0000000001.

## 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 proportion response of a new treatment to that of the current treatment. The current response proportion is 0.53. The health community will be interested in the new treatment if the response rate is increased to 0.63. So that the sample size requirements for different effect sizes can be compared, it is also of interest to compute the sample size at response rates of 0.60, 0.65, 0.70, and 0.75.

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.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
P1 .....	<b>0.53</b>
P2 .....	<b>0.60, 0.63, 0.65, 0.70, 0.75</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>Ignored</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Use Continuity Correction .....	<b>Checked</b>

### Sequential Tab

Number of Looks .....	<b>4</b>
Spending Function .....	<b>O'Brien-Fleming</b>
Times .....	<b>Equally Spaced</b>
Max Time .....	<b>2</b>

### 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 Proportions. Continuity Correction Applied.

Power	N1	N2	Alpha	Beta	P1	P2
0.900168	1102	1102	0.050000	0.099832	0.53	0.60
0.900930	542	542	0.050000	0.099070	0.53	0.63
0.900408	376	376	0.050000	0.099592	0.53	0.65
0.901093	187	187	0.050000	0.098907	0.53	0.70
0.903126	111	111	0.050000	0.096874	0.53	0.75

#### 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.

P1 is the value of both proportions under the null hypothesis.

P2 is the proportion in group two under the alternative hypothesis.

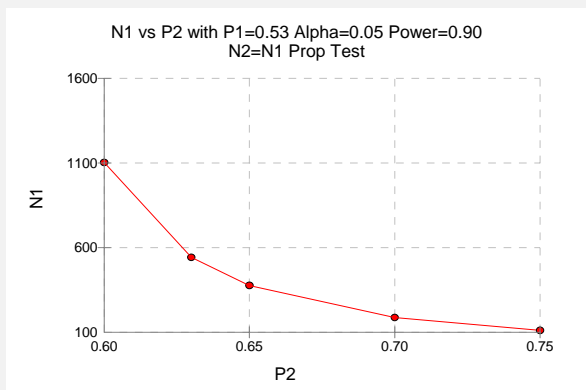
#### Summary Statements

Sample sizes of 1102 and 1102 achieve 90% power to detect a difference of 0.07 between the group proportions of 0.53 and 0.60 at a significance level (alpha) of 0.0500 using a two-sided z-test with continuity correction. 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 542 participants in each arm of the study are required to meet the 90% power requirement when the proportion is 0.63.

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 when the detectable proportion in group two is less than 0.63.

## Details Section

Details when Spending = O'Brien-Fleming, N1 = 542, N2 = 542, P1 = 0.53, P2 = 0.63, Continuity Correction.

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	0.003525	0.003525
2	1.00	-2.96311	2.96311	0.003045	0.003036	0.003051	0.255573	0.259098
3	1.50	-2.35902	2.35902	0.018323	0.016248	0.019299	0.427801	0.686899
4	2.00	-2.01406	2.01406	0.044003	0.030701	0.050000	0.214031	0.900930
Drift	3.27640							

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 0.018323. However, the entry is 0.016248. 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.

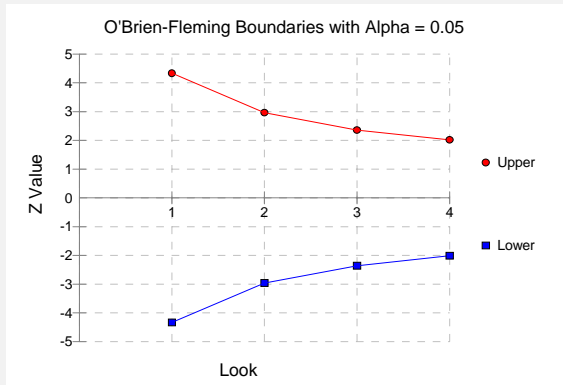
### Total Power

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

Continuing the scenario began in Example1, the researcher wishes to calculate the power of the design at sample sizes 200, 400, 600, 800, 1000. Testing will be done at the 0.01, 0.05, 0.10 significance levels and the overall power will be set to 0.10. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
P1 .....	<b>0.53</b>
P2 .....	<b>0.63</b>
Alpha .....	<b>0.01, 0.05, 0.10</b>
Beta .....	<b>Ignored</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>200 to 1000 by 200</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Use Continuity Correction .....	<b>Checked</b>

### Sequential Tab

Number of Looks .....	<b>4</b>
Spending Function .....	<b>O'Brien-Fleming</b>
Times .....	<b>Equally Spaced</b>
Max Time .....	<b>2</b>

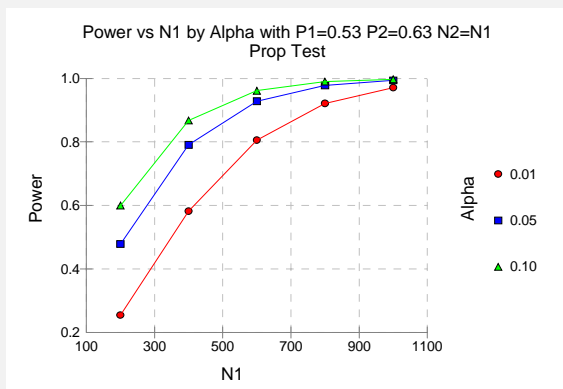
### Annotated Output

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

## Numeric Results

**Numeric Results for Two-Sided Test of Proportions. Continuity Correction Applied.**

Power	N1	N2	Alpha	Beta	P1	P2
0.254797	200	200	0.010000	0.745203	0.53	0.63
0.581920	400	400	0.010000	0.418080	0.53	0.63
0.805679	600	600	0.010000	0.194321	0.53	0.63
0.920951	800	800	0.010000	0.079049	0.53	0.63
0.970898	1000	1000	0.010000	0.029102	0.53	0.63
0.478378	200	200	0.050000	0.521622	0.53	0.63
0.790815	400	400	0.050000	0.209185	0.53	0.63
0.928267	600	600	0.050000	0.071733	0.53	0.63
0.977859	800	800	0.050000	0.022141	0.53	0.63
0.993673	1000	1000	0.050000	0.006327	0.53	0.63
0.599827	200	200	0.100000	0.400173	0.53	0.63
0.867330	400	400	0.100000	0.132670	0.53	0.63
0.961331	600	600	0.100000	0.038669	0.53	0.63
0.989659	800	800	0.100000	0.010341	0.53	0.63
0.997396	1000	1000	0.100000	0.002604	0.53	0.63



These data show the power for various sample sizes and alphas. It is interesting to note that once the sample size is greater than 700, 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 500, 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
P1 .....	<b>0.53</b>
P2 .....	<b>.63</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>Ignored</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>500</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Use Continuity Correction .....	<b>Unchecked</b>

### Sequential Tab

Number of Looks .....	<b>1 (Also run with 2, 3, 4, 6, 8, 10, and 20)</b>
Spending Function .....	<b>O'Brien-Fleming</b>
Times .....	<b>Equally Spaced</b>
Max Time .....	<b>2</b>

### Numeric Results

**Numeric Results for Two-Sided Hypothesis Test of Proportions**

Power	N1	N2	Alpha	Beta	P1	P2	Looks
0.893174	500	500	0.050000	0.106826	0.53	0.63	1
0.892118	500	500	0.050000	0.107882	0.53	0.63	2
0.889620	500	500	0.050000	0.110380	0.53	0.63	3
0.887691	500	500	0.050000	0.112309	0.53	0.63	4
0.885125	500	500	0.050000	0.114875	0.53	0.63	6
0.883535	500	500	0.050000	0.116465	0.53	0.63	8
0.882456	500	500	0.050001	0.117544	0.53	0.63	10
0.879929	500	500	0.050001	0.120071	0.53	0.63	20

This analysis shows how little the number of looks impacts the power of the design. The power of a study with no interim looks is 0.893174. When twenty interim looks are made, the power falls to 0.879929—a very small change.

## Example 4 - Studying a Boundary Set

Continuing with the previous examples, suppose that you are presented with a set of boundaries and want to find the quality of the design (as measured by alpha and power). This is easy to do with *PASS*. Suppose that the analysis is to be run with five interim looks at equally spaced time points. The upper boundaries to be studied are 3.5, 3.5, 3.0, 2.5, 2.0. The lower boundaries are symmetric. The analysis would be run as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
P1 .....	<b>0.53</b>
P2 .....	<b>0.63</b>
Alpha .....	<b>0.05 (will be calculated from boundaries)</b>
Beta .....	<b>Ignored</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>500</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Use Continuity Correction .....	<b>Checked</b>
<b>Sequential Tab</b>	
Number of Looks .....	<b>5</b>
Spending Function .....	<b>User Supplied</b>
Times .....	<b>Equally Spaced</b>
Lower Boundaries .....	<b>Symmetric</b>
Upper Boundaries .....	<b>3.5, 3.5, 3.0, 2.5, 2.0</b>
Max Time .....	<b>2</b>



## Numeric Results

### Numeric Results for Two-Sided Hypothesis Test of Proportions

Power	N1	N2	Alpha	Beta	P1	P2
0.887792	500	500	0.048157	0.112208	0.53	0.63

### Details when Spending = User Supplied, N1 = 500, N2 = 500, P1 = 0.53, P2 = 0.63

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	0.018094	0.018094
2	0.80	-3.50000	3.50000	0.000465	0.000408	0.000874	0.054010	0.072105
3	1.20	-3.00000	3.00000	0.002700	0.002410	0.003284	0.219472	0.291577
4	1.60	-2.50000	2.50000	0.012419	0.010331	0.013615	0.335232	0.626809
5	2.00	-2.00000	2.00000	0.045500	0.034542	0.048157	0.248848	0.875657
Drift	3.14209							

The power for this design is about 0.88. This value depends on both the boundaries and the sample size. The alpha level is about 0.048. This value only depends on the boundaries.

## Example5 - Validation Using O'Brien-Fleming Boundaries

Reboussin (1992) presents an example for binomial distributed data for a design with two-sided O'Brien-Fleming boundaries, looks = 5, alpha = 0.05, beta = 0.10, P1 = 0.1100, P2 = 0.0825. They compute a drift of 3.28 and a sample size of 2381.78 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N1</b>
P1 .....	<b>0.1100</b>
P2 .....	<b>0.0825</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
N1 .....	<b>Ignored</b>
N2 .....	<b>Use R</b>
R .....	<b>1.0</b>
Use Continuity Correction .....	<b>Not checked</b>

## Sequential Tab

Number of Looks.....5  
 Spending Function .....O'Brien-Fleming  
 Times .....Equally Spaced  
 Max Time .....1

## Numeric Results

### Numeric Results for Two-Sided Hypothesis Test of Proportions

Power	N1	N2	Alpha	Beta	P1	P2
0.900105	2474	2474	0.050000	0.099895	0.1100	0.0825

Details when Spending = O'Brien-Fleming, N1 = 2468, N2 = 2468, P1 = 0.1100, P2 = 0.0825

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	0.000324	0.000324
2	0.40	-3.35695	3.35695	0.000788	0.000787	0.000788	0.099454	0.099778
3	0.60	-2.68026	2.68026	0.007357	0.006828	0.007616	0.346699	0.446477
4	0.80	-2.28979	2.28979	0.022034	0.016807	0.024424	0.299644	0.746120
5	1.00	-2.03100	2.03100	0.042255	0.025576	0.050000	0.153985	0.900105
Drift	3.27939							

The difference in the sample sizes (2468 versus 2382) is due to rounding errors in the Reboussin article. Reboussin rounds from four-digits to three-digits, which caused a large difference. *PASS* uses more accurate routines.

To see that the results are equal to within rounding error, we will compute the sample size using Reboussin's results, but with more decimal places in the intermediate steps. They had

$$n_K = \frac{2(0.096)(0.904)(3.28)^2}{(0.028)^2} = 2381.78$$

When we compute this without rounding, we get

$$n_K = \frac{2(0.09625)(0.90375)(3.27939)^2}{(0.0275)^2} = 2474.00$$

The 2474 is what *PASS* gets.

## Chapter 225

# Comparing Proportions in Stratified Designs (Mantel-Haenszel)

## Introduction

In a stratified design, the subjects are selected from two or more strata which are formed from important covariates such as gender, income level, or marital status. The number of subjects in each of the two groups in each strata is set (fixed) by the design. A separate 2-by-2 table is formed for each stratum. Although response rates may vary among strata, hypotheses about the overall odds ratio can be tested the Cochran-Mantel-Haenszel test. This module allows you to determine power and sample size for such a study.

## Technical Details

This procedure is based on the results of Woolson, Bean, and Rojas (1986) which were extended to include a continuity correction by Nam (1992). For more details, consult those articles or chapter 4 in Lachin (2000). We will now briefly summarize these results.

Suppose you are interested in comparing the disease response rates of two groups (treatment and control). Further suppose that response rate is known to be related to another covariate (such as age, race, or gender). It is often desirable to remove the covariate's impact from the comparison of the two proportions. This is accomplished by stratifying on the covariate and forming hypotheses about a common odds ratio across all strata. Data from such a stratified design may be analyzed by the Cochran-Mantel-Haenszel test.

There are two versions of the Cochran-Mantel-Haenszel test: one that is continuity corrected and one that is not. The continuity-corrected test is more commonly used.

The computation of the test statistic is as follows. Suppose there are  $J$  strata. The result of each 2-by-2 table may be summarized as follows.

	<u>Groups</u>		
	<b>Group 1</b>	<b>Group 2</b>	
<u>Response</u>	<u>Treatment</u>	<u>Control</u>	<b>Total</b>
<b>Yes</b>	$x_{1j}$	$x_{2j}$	$x_{.j}$
<b>No</b>	$n_{1j} - x_{1j}$	$n_{2j} - x_{2j}$	$N_j - x_{.j}$
<b>Total</b>	$n_{1j}$	$n_{2j}$	$N_j$

where  $j = 1, 2, \dots, J$  and  $N = \sum_{j=1}^J N_j$ .

The parameters of interest are the success proportions  $p_{1j}$  and  $p_{2j}$ . These parameters are estimated by

$$\hat{p}_{1j} = \frac{x_{1j}}{n_{1j}} \text{ and } \hat{p}_{2j} = \frac{x_{2j}}{n_{2j}}$$

The odds of response in each of the two groups in each strata is given by

$$o_{1j} = \frac{p_{1j}}{1 - p_{1j}} \text{ and } o_{2j} = \frac{p_{2j}}{1 - p_{2j}}$$

The strata odds ratio  $\psi_j$  is calculated using the equation

$$\begin{aligned} \psi_j &= \frac{o_{1j}}{o_{2j}} \\ &= \frac{\left( \frac{p_{1j}}{1 - p_{1j}} \right)}{\left( \frac{p_{2j}}{1 - p_{2j}} \right)} \end{aligned}$$

In the sequel, it is assumed that the strata odds ratios are all equal. That is, it is assumed that  $\psi_1 = \psi_2 = \dots = \psi_J = \psi$ . Solving this relationship for  $p_{1j}$  in terms of  $\psi$  and  $p_{2j}$  gives

$$p_{1j} = \frac{\psi p_{2j}}{1 - p_{2j} + \psi p_{2j}}$$

If values for the odds ratio under the null hypothesis ( $\psi_0$ ), under the alternative hypothesis ( $\psi_1$ ), and  $p_{2j}$  are specified, values for  $p_{1j}$  under the null hypothesis ( $p_{1j0}$ ) and the alternative hypothesis ( $p_{1j1}$ ) can be calculated as follows

$$p_{1j0} = \frac{\psi_0 p_{2j}}{1 - p_{2j} + \psi_0 p_{2j}}, \quad j = 1, 2, \dots, J$$

$$p_{1j1} = \frac{\psi_1 p_{2j}}{1 - p_{2j} + \psi_1 p_{2j}}, \quad j = 1, 2, \dots, J$$

Assuming a common odds ratio across all strata of  $\psi$  (that is, assuming  $\psi_1 = \psi_2 = \dots = \psi_J = \psi$ ), hypotheses of the form  $H_0: \psi \leq \psi_0$  versus  $H_1: \psi > \psi_0$  may be tested using Cochran's  $U$  statistic (Woolson et al. 1986, page 928)

$$U_G = \sum_{j=1}^J w_j \left\{ (\hat{p}_{1j} - \hat{p}_{2j}) - (p_{1j0} - p_{2j}) \right\}, \quad \text{where } w_j = \frac{n_{1j}n_{2j}}{N_j}$$

Note that when  $\psi_0 = 1$ ,  $U_G$  reduces to

$$U_0 = \sum_{j=1}^J w_j (\hat{p}_{1j} - \hat{p}_{2j}).$$

The value  $U_0$  is commonly used to form the Cochran-Mantel-Haenszel statistic.  $U_G$  is an extension of this statistic which allows  $\psi_0 \neq 1$ .

The calculation of the asymptotically normal test statistic,  $z_c$ , may or may not include a continuity correction factor depending on whether the parameter  $cc$  is set to 1/2 or 0. The formula for  $z_{CMH}$  is

$$z_{CMH} = \frac{U_G - cc}{\sqrt{v_0(U_G)}}$$

where

$$v_0(U_G) = \begin{cases} \sum_{j=1}^J w_j^2 \left\{ \frac{\hat{p}_{1j}(1 - \hat{p}_{1j})}{n_{1j}} + \frac{\hat{p}_{2j}(1 - \hat{p}_{2j})}{n_{2j}} \right\} & \text{if } \psi_0 \neq 1 \\ \sum_{j=1}^J w_j \hat{p}_j (1 - \hat{p}_j) & \text{if } \psi_0 = 1 \end{cases}$$

$$\hat{p}_j = \frac{x_{.j}}{N_j}$$

The name *Cochran-Mantel-Haenszel test* actually refers to two tests: the Cochran test and the Mantel-Haenszel test. The difference between these test is that Cochran's test uses  $v_0(U_G)$  to estimate the unconditional variance assuming that the group sample sizes are fixed, while the Mantel-Haenszel test replaces  $v_0(U_G)$  with an estimate of the conditional variance of  $U$  assuming that both row and column marginals are fixed. Asymptotically the two variances are equivalent, so the test is often called the Cochran-Mantel-Haenszel statistic.

## Power Calculations

The asymptotic power of  $z_{CMH}$  for testing a one-sided hypothesis of the form  $H_0: \psi \leq \psi_0$  versus  $H_1: \psi > \psi_0$  is

$$Power = 1 - \Phi \left( \frac{z_{1-\alpha} \sqrt{V_0(U_G)} - E(U_G) + cc}{\sqrt{V_1(U_G)}} \right)$$

where

$$E(U_G) = \sum_{j=1}^J w_j \{ (p_{1j1} - p_{2j}) - (p_{1j0} - p_{2j}) \}$$

$$V_0(U_G) = \begin{cases} \sum_{j=1}^J w_j^2 \left\{ \frac{p_{1j0}(1 - p_{1j0})}{n_{1j}} + \frac{p_{2j}(1 - p_{2j})}{n_{2j}} \right\} & \text{if } \psi_0 \neq 1 \\ \sum_{j=1}^J w_j \bar{p}_j (1 - \bar{p}_j) & \text{if } \psi_0 = 1 \end{cases}$$

$$\bar{p}_j = p_{1j1} \left( \frac{n_{1j}}{N_j} \right) + p_{2j} \left( \frac{n_{2j}}{N_j} \right)$$

$$V_1(U_G) = \sum_{j=1}^J w_j^2 \left\{ \frac{p_{1j1}(1 - p_{1j1})}{n_{1j}} + \frac{p_{2j}(1 - p_{2j})}{n_{2j}} \right\}$$

Note that Woolson et al. (1986) and Nam (1992) give results for the usual case when  $\psi_0 = 1$ . The above results are our extension to the important case when  $\psi_0 \neq 1$ . We could not find published results for this case, so we have made this extension. When published results become available, we will adopt those results. If you have  $\psi_0 \neq 1$ , you must use  $U_G$ , rather than  $U_0$ , in the calculation of the test statistic.

Similar calculations may also be made for testing the other one-sided hypothesis  $H_0: \psi \geq \psi_0$  versus  $H_1: \psi < \psi_0$  and the two-sided hypothesis  $H_0: \psi = \psi_0$  versus  $H_1: \psi \neq \psi_0$ .

# 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, refer to the Procedure Templates chapter.

## Data Tab

The Data tab contains most of the parameters and options of interest for this procedure.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *OR1*, *Alpha*, *Beta*, or *N*. In most cases, you will select either *Beta* or *N*.

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.

### OR1 (Odds Ratio|H1)

This option specifies the odds ratio of the two proportions P1 and P2 at which the power is to be computed. This odds ratio is used to specify the size of the difference between the two proportions at which the power is calculated.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*.

Odds ratios must greater than zero.

### OR0 (Odds Ratio|H0)

Specify the odds ratio under the null hypothesis, H0. For each strata, this value is used with the value of Pr(Success) to calculate the probability of obtaining a success in group one (the treatment group) assuming the null hypothesis. In the standard Cochran-Mantel-Haenszel test, this value is assumed to be (and should be entered as) one. If you enter a value other than one, your data analysis should use the more general test statistic.

Note that OR0 must be greater than zero and cannot be equal to OR1.

### Alternative Hypothesis (H1)

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided (H1:  $OR1 < OR0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: OR1 < OR0$ .

*One-Sided (H1:  $OR1 > OR0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: OR1 > OR0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: OR1 <> OR0$ . Here ' $<>$ ' means 'is not equal to' or 'is less than or greater than'.

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

## 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 proportions when in fact they 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. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

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

## 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 proportions when in fact they are different.

Values must be between zero and one. Historically, the value of 0.20 was often 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  $1 - \text{beta}$ . 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 power values of 0.95, 0.90, and 0.80, respectively.

## M (Sample Size Multiplier)

M and the values of R1 and R2 are used to calculate the group sample sizes within each strata using the formulas  $N1 = M \times R1$  and  $N2 = M \times R2$ . The total sample size, N, is found by summing N1 and N2 across all strata. Note that fractional values for N, N1, and N2 will usually result. In practice these values are rounded up to the next integer value.

One or more values, separated by blanks or commas, may be entered. A separate analysis is performed for each value.

### Using M as the Group Size

To use M as the sample size in each group, the values of R1 and R2 must each be set to one.

### Using M as the Strata Size:

To use M as the sample size in each strata, the values of R1 and R2 must sum to one within each strata. For example, suppose  $M = 30$  and  $R1 = R2 = 0.5$ . The values of N1 and N2, the group sample sizes within a stratum, will be  $0.5 \times 30 = 15$ . Thus, the total sample size within the strata is  $15 + 15 = 30$ .

### Using M as Total Sample Size:

To use M as the total sample size across all strata, the values of R1 and R2 must sum to one across all values. Note that the resulting value of N may not exactly equal M because of rounding.

For example, suppose there are three strata with  $R1 = 0.1, 0.2, \text{ and } 0.2$  and  $R2 = 0.1, 0.3, \text{ and } 0.1$ . (Note that these values sum to one.) If M were 100, then the values of N1 would be 10, 20, and 20 and the values of N2 would be 10, 30, and 10. These sum to 100, the value of M.



## Strata

This option specifies the number of strata specified on this line. Usually, you will enter a '1' to specify a single stratum, or you will enter a '0' to ignore this line. However, this option lets you specify several strata at once.

The total number of strata is equal to the sum of these values.

## $R1 = N1 / M, R2 = N2 / M$

R1 and R2 are used to obtain the sample sizes in groups 1 (treatment) and 2 (control) within a strata using the formulas  $N1 = R1 \times M$  and  $N2 = R2 \times M$ . The only limitation on R1 and R2 is that they are positive (non-zero) values. See the comments under M for more information.

Note that only a single value may be entered for this parameter—you cannot enter several values.

## Pr(Success)

This is the baseline probability of a successful response. This value is used with OR1 to calculate the probability of a success in group 1 (the treatment or numerator group).

Since this value is a probability, it must be between zero and one.

Note that only one value may be entered here.

## Continuity Correction

Specify whether to use the Continuity Correction. When selected, a continuity correction is made that is recommend by Fleiss et al. (2003) to make the alpha and beta values achieved by the test more accurate.

## Example1 - Finding Power

Nam (1992) discusses a case-control study investigating the possible association between chlorinated water and colon cancer among males in Iowa. Since age is known to affect colon cancer rates, the population is stratified into four age groups with weights of 10%, 40%, 35%, and 15%. An equal number of cases and controls will be selected in each age-group. Prior studies had shown the probability of chlorinated water exposure among non-cancer subjects was 0.75, 0.70, 0.65, and 0.60, respectively, among the four age groups. The significance level is set to 0.05. The investigators want to consider various total sample sizes from 50 to 500. They also want to consider odds ratios of 2 and 3.

### 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Hypothesis .....	<b>One-Sided (H1:OR1&gt;OR0)</b>
OR0.....	<b>1</b>
OR1.....	<b>2 3</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
M.....	<b>50 to 500 by 50</b>
Continuity Correction .....	<b>Checked</b>
Strata(1) .....	<b>1</b>
R1(1) .....	<b>0.05 (half of 10%)</b>
R2(1) .....	<b>R1</b>
Pr(Success)(1) .....	<b>0.75</b>
Strata(2) .....	<b>1</b>
R1(2) .....	<b>0.20 (half of 40%)</b>
R2(2) .....	<b>R1</b>
Pr(Success)(2) .....	<b>0.70</b>
Strata(3) .....	<b>1</b>
R1(3) .....	<b>0.175 (half of 35%)</b>
R2(3) .....	<b>R1</b>
Pr(Success)(3) .....	<b>0.65</b>
Strata(4) .....	<b>1</b>
R1(4) .....	<b>0.075 (half of 15%)</b>
R2(4) .....	<b>R1</b>
Pr(Success)(4) .....	<b>0.60</b>

### Annotated Output

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

## Numeric Results

**Numeric Results of Cochran-Mantel-Haenszel Test of an Odds Ratio**  
**H0:  $OR_1=OR_0$ . H1:  $OR_1>OR_0$ . Test: Continuity-Corrected Z-Test.**

	Total Sample Size	Sample Size Multiplier	Sample Size of Group 1	Sample Size of Group 2	H0 Odds Ratio (OR0)	Actual Odds Ratio (OR1)	Signif. Level Alpha	Beta
Power	(N)	(M)	(N1)	(N2)				
0.1783	50	50.000	25	25	1.000	2.000	0.0500	0.8217
0.3505	100	100.000	50	50	1.000	2.000	0.0500	0.6495
0.4992	150	150.000	75	75	1.000	2.000	0.0500	0.5008
0.6215	200	200.000	100	100	1.000	2.000	0.0500	0.3785
0.7186	250	250.000	125	125	1.000	2.000	0.0500	0.2814
0.7937	300	300.000	150	150	1.000	2.000	0.0500	0.2063
0.8506	350	350.000	175	175	1.000	2.000	0.0500	0.1494
0.8929	400	400.000	200	200	1.000	2.000	0.0500	0.1071
0.9239	450	450.000	225	225	1.000	2.000	0.0500	0.0761
0.9464	500	500.000	250	250	1.000	2.000	0.0500	0.0536
0.3356	50	50.000	25	25	1.000	3.000	0.0500	0.6644
0.6337	100	100.000	50	50	1.000	3.000	0.0500	0.3663
0.8151	150	150.000	75	75	1.000	3.000	0.0500	0.1849
0.9121	200	200.000	100	100	1.000	3.000	0.0500	0.0879
0.9601	250	250.000	125	125	1.000	3.000	0.0500	0.0399
0.9825	300	300.000	150	150	1.000	3.000	0.0500	0.0175
0.9925	350	350.000	175	175	1.000	3.000	0.0500	0.0075
0.9969	400	400.000	200	200	1.000	3.000	0.0500	0.0031
0.9987	450	450.000	225	225	1.000	3.000	0.0500	0.0013
0.9995	500	500.000	250	250	1.000	3.000	0.0500	0.0005

### Report Definitions

'Power' is the probability of rejecting a false null hypothesis. It should be close to one.

'N' is the total sample size summed across all groups and strata.

'M' is the factor by which the values of R1 and R2 are multiplied.

'N1 and N2' are the sample sizes from groups 1 and 2 summed across all strata.

'OR0' is the odds ratio  $[P1/(1-P1)] / [P2/(1-P2)]$  assuming the null hypothesis (H0).

'OR1' is the value of the odds ratio at which the power is computed.

'Alpha' is the probability of rejecting a true null hypothesis.

'Beta' is the probability of accepting a false null hypothesis.

In a treatment vs. control design, the treatment group is 1 and the control group is 2.

### Summary Statements

A stratified design, which divides the sample among 4 strata, is analyzed using the one-sided, Cochran-Mantel-Haenszel test. Sample sizes, summed across all strata, of 25 in group 1 (treatment group) and 25 in group 2 (control group) achieve 18% power to reject the odds ratio set by the null hypothesis of 1.000 when the odds ratio is actually 2.000. The significance level of the test was set at 0.0500.

## Sample Sizes: N, N1, and N2

The value of N is the sum of N1 and N2. The values of N1 and N2 are found by summing the individual strata-group sample sizes. These are found by multiplying R1 and R2 by M.

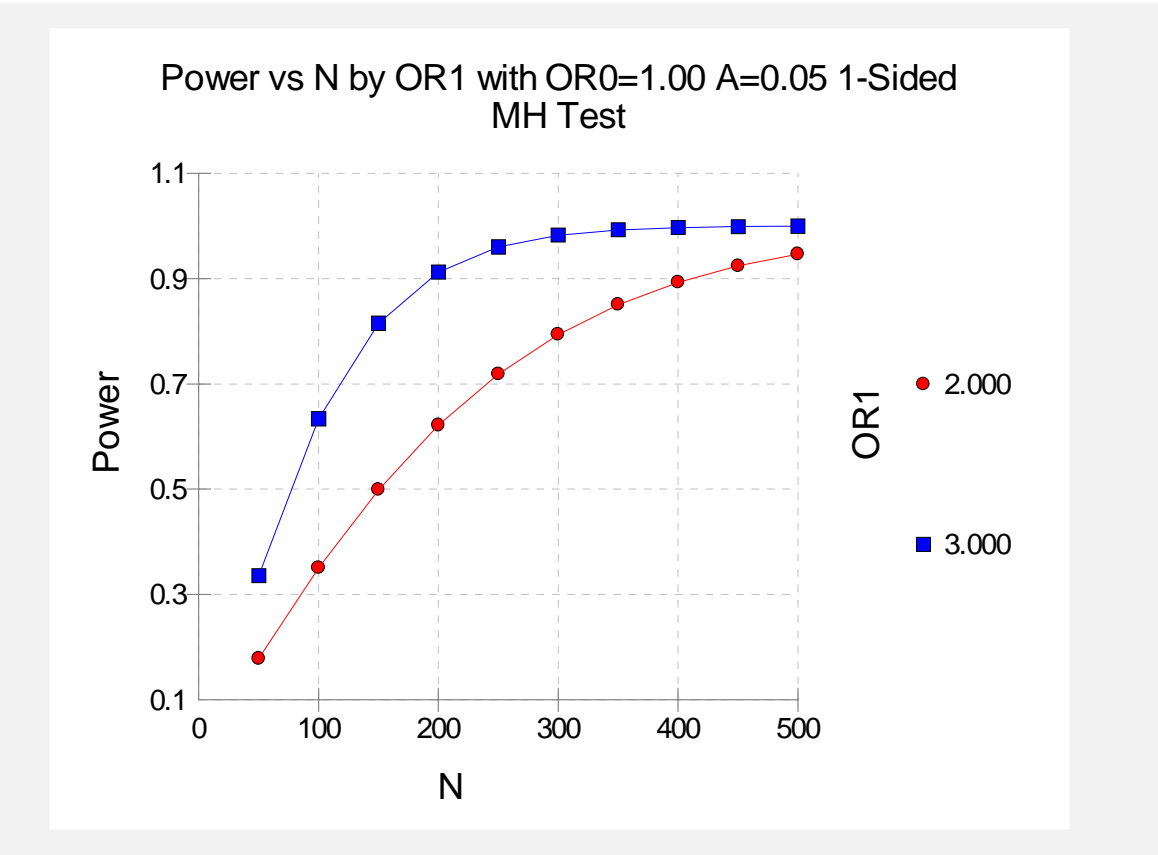
Note that this multiplication will usually result in fractional sample sizes across the strata. As a practical matter, we recommend rounding each fractional value up to the next integer when implementing a given design.

# Strata-Detail Report

Strata-Detail Report						
Number of Strata	Proportion of Total Sample in each Strata	Proportion of this Strata in Group 1	Proportion of this Strata in Group 2	Group 1 Multiplier (R1)	Group 2 Multiplier (R2)	Strata Probability of Success
1	0.1000	0.5000	0.5000	0.050	0.050	0.7500
1	0.4000	0.5000	0.5000	0.200	0.200	0.7000
1	0.3500	0.5000	0.5000	0.175	0.175	0.6500
1	0.1500	0.5000	0.5000	0.075	0.075	0.6000

This report shows the values of the individual, strata-level parameters that were used. These parameters are the same for all rows of the Numerical Results Report (shown above), so they are only displayed once.

## Plots Section



The values from the Numerical Results Report are displayed in this scatter plot. This chart provides a quick view of the power that is achieved for various sample sizes.

## Example2 - Validation Using Nam

To validate the procedure, we will compare *PASS*'s results to those on page 392 of Nam (1992). Most of the settings in this example are the same as those of Example 1, except that the power is 90% and the odds ratio is 3. Nam (1992) found the necessary sample sizes to be 192 for the corrected test and 171 for the uncorrected test.

### 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 either template Example2a or Example2b by clicking the Template tab and loading the desired template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N (Sample Size)</b>
Alternative Hypothesis .....	<b>One-Sided (H1:OR1&gt;OR0)</b>
OR0 .....	<b>1</b>
OR1 .....	<b>3</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10</b>
M .....	<i>Ignored since this is the Find setting</i>
Continuity Correction .....	<b>Checked</b>
Strata(1) .....	<b>1</b>
R1(1) .....	<b>0.05 (half of 10%)</b>
R2(1) .....	<b>R1</b>
Pr(Success)(1) .....	<b>0.75</b>
Strata(2) .....	<b>1</b>
R1(2) .....	<b>0.20 (half of 40%)</b>
R2(2) .....	<b>R1</b>
Pr(Success)(2) .....	<b>0.70</b>
Strata(3) .....	<b>1</b>
R1(3) .....	<b>0.175 (half of 35%)</b>
R2(3) .....	<b>R1</b>
Pr(Success)(3) .....	<b>0.65</b>
Strata(4) .....	<b>1</b>
R1(4) .....	<b>0.075 (half of 15%)</b>
R2(4) .....	<b>R1</b>
Pr(Success)(4) .....	<b>0.60</b>

### Annotated Output

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

## Numeric Results

Numeric Results of Cochran-Mantel-Haenszel Test of an Odds Ratio

H0:  $OR_1=OR_0$ . H1:  $OR_1>OR_0$ . Test: Continuity-Corrected Z-Test.

	Total Sample Size (N)	Sample Size Multiplier (M)	Sample Size of Group 1 (N1)	Sample Size of Group 2 (N2)	H0 Odds Ratio (OR0)	Actual Odds Ratio (OR1)	Signif. Level Alpha	Beta	
Power	0.9000	192	191.538	96	96	1.000	3.000	0.0500	0.1000

The value of 192 agrees exactly with that of Nam (1992).

If you uncheck the Continuity Correction option and rerun the analysis, you will get the following results.

Numeric Results of Cochran-Mantel-Haenszel Test of an Odds Ratio

H0:  $OR_1=OR_0$ . H1:  $OR_1>OR_0$ . Test: Uncorrected Z-Test.

	Total Sample Size (N)	Sample Size Multiplier (M)	Sample Size of Group 1 (N1)	Sample Size of Group 2 (N2)	H0 Odds Ratio (OR0)	Actual Odds Ratio (OR1)	Signif. Level Alpha	Beta	
Power	0.9000	171	170.741	85	85	1.000	3.000	0.0500	0.1000

The value of 171 agrees exactly with that of Nam (1992).

## Example3 - Finding Power of a Completed Experiment

Suppose you want to find the power for a completed experiment in which the individual strata sample sizes are known. In this example there are three strata with success probabilities 0.72, 0.66, and 0.69. The sample sizes for the treatment group in each stratum are 102, 113, and 97. The sample sizes for the control group in each stratum are 98, 110, and 114. The experiment was designed to detect an odds ratio of at least 1.5 with alpha equal to 0.05 for a one-sided test.

To calculate the power in this situation, we set M to 1 and enter the sample sizes directly into R1 and R2.

### 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Alternative Hypothesis .....	<b>One-Sided (H1:OR1&gt;OR0)</b>
OR0 .....	<b>1</b>
OR1 .....	<b>1.5</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>
M .....	<b>1</b>
Continuity Correction.....	<b>Checked</b>
Strata(1) .....	<b>1</b>
R1(1) .....	<b>102</b>
R2(1) .....	<b>98</b>
Pr(Success)(1) .....	<b>0.72</b>
Strata(2) .....	<b>1</b>
R1(2) .....	<b>113</b>
R2(2) .....	<b>110</b>
Pr(Success)(2) .....	<b>0.66</b>
Strata(3) .....	<b>1</b>
R1(3) .....	<b>97</b>
R2(3) .....	<b>114</b>
Pr(Success)(3) .....	<b>0.69</b>

### Annotated Output

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

## Numeric Results

Numeric Results of Cochran-Mantel-Haenszel Test of an Odds Ratio  
H0:  $OR_1=OR_0$ . H1:  $OR_1>OR_0$ . Test: Continuity-Corrected Z-Test.

	Total Sample Size (N)	Sample Size Multiplier (M)	Sample Size of Group 1 (N1)	Sample Size of Group 2 (N2)	H0 Odds Ratio (OR0)	Actual Odds Ratio (OR1)	Signif. Level Alpha	Beta	
Power	0.6980	634	1.000	312	322	1.000	1.500	0.0500	0.3020

The power to detect an odds ratio of 1.5 is only 0.6980 in this experiment.



## Chapter 230

# Comparing Proportions in Cluster-Randomized Trials

## Introduction

A *cluster (group) randomized design* is one in which whole units, or clusters, of subjects are randomized to the groups rather than the individual subjects in those clusters. However, the conclusions of the study concern individual subjects rather than the clusters. Examples of clusters are families, school classes, neighborhoods, and hospital wards.

Cluster-randomized designs are often adopted when there is a high risk of contamination if cluster members were randomized individually. For example, it may be difficult for an instructor to use two methods of teaching individuals in the same class. The price of randomizing by clusters is a loss of efficiency--the number of subjects needed to obtain a certain level of precision in a cluster-randomized trial is usually much larger than the number needed when the subjects are randomized individually. Hence, the standard methods of sample size estimation cannot be used.

## Three Procedures Documented Here

There are three procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, or ratios of proportions. Each of these options is listed separately on the menus.

## Technical Details

Our formulation comes from Donner and Klar (2000). Denote a binary observation by  $Y_{gkm}$  where  $g = 1$  or  $2$  is the group,  $k = 1, 2, \dots, K_g$  is a cluster within group  $g$ , and  $m = 1, 2, \dots, M_g$  is an individual in cluster  $k$  of group  $g$ . The results that follow assume an equal number of individuals per cluster. When the number of subjects from cluster to cluster are about the same, the power and sample size values should be fairly accurate. In these cases, the average number of subjects per cluster can be used.

The statistical hypothesis that is tested concerns the difference between the two group proportions,  $p_1$  and  $p_2$ . When necessary, we assume that group 1 is the treatment group and group 2 is the control group. With a simple modification, all of the large-sample sample size formulas that are listed in the module for testing two proportions can be used here. When the individual subjects are randomly assigned to one of the two groups, the variance of the sample proportion is

$$\sigma_{S,g}^2 = \frac{p_g(1-p_g)}{n_g}$$

When the randomization is by clusters of subjects, the variance of the sample proportion is

$$\begin{aligned}\sigma_{C,g}^2 &= \frac{p_g(1-p_g)(1+(m_g-1)\rho)}{k_g m_g} \\ &= \sigma_{S,g}^2 [1+(m_g-1)\rho] \\ &= F_{g,\rho} \sigma_{S,g}^2\end{aligned}$$

The factor  $[1+(m_g-1)\rho]$  is called the *inflation factor*. The Greek letter  $\rho$  is used to represent the *intraclass correlation coefficient (ICC)*. This correlation may be thought of as the simple correlation between any two subjects within the same cluster. If we stipulate that  $\rho$  is positive, it may also be interpreted as the proportion of total variability that is attributable to differences between clusters. This value is critical to the sample size calculation.

All of the asymptotic formulas that were used in comparing two proportions may be used with cluster-randomized designs as well, as long as an adjustment is made for the inflation factor. The basic form of the z-test becomes

$$z = \frac{|\hat{D} - \delta_0|}{\hat{\sigma}_{\hat{D}}(\delta_0)}$$

where

$$\begin{aligned}\hat{D} &= \hat{p}_1 - \hat{p}_2 \\ \delta_0 &= p_1 - p_2 | H_0 \\ \hat{\sigma}_{\hat{D}}(\delta_0) &= \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)F_{1,\rho}}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)F_{2,\rho}}{n_2}}\end{aligned}$$

The quantities  $\tilde{p}_1$  and  $\tilde{p}_2$  are the maximum likelihood estimates constrained by  $\tilde{p}_1 - \tilde{p}_2 = \delta_0$ .

## Power Calculations

A large sample approximation may be used that is most accurate when the values of  $n_1$  and  $n_2$  are large. The large approximation is made by replacing the values of  $\hat{p}_1$  and  $\hat{p}_2$  in the  $z$  statistic with the corresponding values of  $p_1$  and  $p_2$  under the alternative hypothesis, and then computing the results based on the normal distribution.

Note that in this case, exact calculations are not possible.

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers three procedures, each of which has different options. This section documents options that are common to all three procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *P1.1*, *Alpha*, *Beta*, *K1*, and *M1*. Under most situations, you will select either *Beta* or *K1*.

Select *K1* 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.

### Test Statistic

Specify which test statistic is used in searching and reporting. We recommend the likelihood score test.

## P2 (Control Group Proportion)

Specify the value of  $p_2$ , the control, baseline, or standard group's proportion. The null hypothesis is that the two proportions differ by a specified amount (See *Specify Group 1 Proportion using below*).

Since  $p_2$  is a proportion, these values must be between zero and one.

You may enter a range of values such as *0.1,0.2,0.3* or *0.1 to 0.9 by 0.1*.

## K1 (Clusters Group 1)

Enter a value (or range of values) for the number of clusters in this group. You may enter a range of values such as *10 to 20 by 2*. The sample size for this group is equal to the number of clusters times the number of subjects per cluster.

## M1 (Items Group 1)

This is the average number of items (subjects) per cluster in group one. This value must be a positive number that is at least one. You can use a list of values such as *100 150 200*.

## K2 (Clusters Group 2)

This is the number of clusters in group two. The sample size for this group is equal to the number of clusters times the number of subjects per cluster. This value must be a positive number.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *K1*, with no blanks. If you want to use *K1* directly, you do not have to premultiply by *1*. For example, all of the following are valid entries: *10 K1 2K1 0.5K1*.

You can use a list of values such as *10 20 30* or *K1 2K1 3K1*.

## M2 (Items Group 2)

This is the average number of items (subjects) per cluster in group two. This value must be at least one.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *M1*, with no blanks. If you want to use *M1* directly, you do not have to premultiply by *1*. For example, all of the following are valid entries: *10 M1 2M1 0.5M1*.

You can use a list of values such as *10 20 30* or *M1 2M1 3M1*.

## 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 proportions when in fact they 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. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

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

## 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 proportions when in fact they are different.

Values must be between zero and one. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.

## ICC (Intraclass Correlation)

Enter a value (or range of values) for the intraclass correlation. This correlation may be thought of as the simple correlation between any two observations in the same cluster. 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 0.05.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions, P1 and P2. P1.0 is the value of the P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated.

### P1.0 (Group 1 Proportion |H0)

This option specifies the value of the group 1 proportion given the null hypothesis. The power calculations assume that P1.0 is the value of P2 under the null hypothesis. In this non-null case, the value of P1.0 is not equal to P2 as it is in the null case.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

Proportions must be between zero and one. They cannot take on the values zero or one.

### P1.1 (Group 1 Proportion |H1)

This is the value of P1 under the alternative hypothesis. It is written P1.1. The power calculations assume that this is the actual value of the proportion.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

Note that values must be between zero and one.

## Alternative Hypothesis (H1)

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided (H1:  $P1 < P2$ )* refers to a one-sided test in which the alternative hypothesis is of the form H1:  $P1 < P2$ .

*One-Sided (H1:  $P1 > P2$ )* refers to a one-sided test in which the alternative hypothesis is of the form H1:  $P1 > P2$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type H1:  $P1 \neq P2 + D0$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $D0$ , and  $D1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### D0 (Difference|H0 = $P1.0 - P2$ )

This option specifies the difference between the two proportions given in the null hypothesis,  $H0$ . This difference is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = P2 + D0$ . Note that  $P1.0$  here means the value of  $P1$  under  $H0$ .

Differences must be between -1 and 1. They cannot take on the values -1, 0, or 1.

The power calculations use  $P1.0$  as the value of the proportion in group 2 (the experimental or treatment group) under the null hypothesis. In the non-null case, the value of  $P1.0$  is not equal to  $P2$  as it is in the null case.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

### D1 (Difference|H1 = $P1.1 - P2$ )

This option specifies the difference between  $P1.1$  and  $P2$ . This difference is used with  $P2$  to calculate the value of  $P1.1$  using the formula:  $P1.1 = D1 + P2$ . Note that  $P1.1$  here means the value of  $P1$  under  $H1$ . Differences must be between -1 and 1. They cannot take on the values -1 or 1.

The power calculations assume that  $P1.1$  is the actual value of the proportion in group 2 (experimental or treatment group).

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

## Alternative Hypothesis (H1)

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided ( $H1:D1 < D0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: D1 < D0$ .

*One-Sided ( $H1:D1 > D0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: D1 > D0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: D1 \neq D0$ . Here ' $\neq$ ' means unequal.

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

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $R0$ , and  $R1$  are given, the values of  $P1.0$  and  $P1.1$  can be calculated.

### **R0 (Ratio|H0 = $P1.0 / P2$ )**

This option specifies the ratio between the group 1 proportion under the null hypothesis,  $P1.0$ , and  $P2$ . This ratio is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = R0 \times P2$ . The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. In this non-null case, the value of  $P1.0$  is not equal to  $P2$  as it is in the null case.

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*.

Ratios must be greater than zero.

### **R1 (Ratio|H1 = $P1.1 / P2$ )**

This option specifies the ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the proportion in group 1 under the alternative hypothesis. This ratio is used with  $P2$  to calculate the value of  $P1.1$  using the formula:  $P1.1 = R1 \times P2$ . The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group).

You may enter a range of values such as *0.5 0.6 0.7 0.8* or *1.25 to 2.0 by 0.25*.

Ratios must be greater than zero. They cannot take on the value of one.

## **Alternative Hypothesis (H1)**

This option specifies whether a one-sided or two-sided hypothesis is analyzed.

*One-Sided ( $H1: R1 < R0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: R1 < R0$ .

*One-Sided ( $H1: R1 > R0$ )* refers to a one-sided test in which the alternative hypothesis is of the form  $H1: R1 > R0$ .

*Two-Sided* refers to a two-sided test in which the alternative hypothesis is of the type  $H1: R1 \neq R0$ . Here ' $\neq$ ' means unequal.

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



# Options Tab

The Options tab contains various limits and options.

## Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

## Zero Count Adjustment Method

Zero cell counts often cause calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option *Add to zero cells only*.

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts. We have found that adding 0.0001 seems to work well.

## Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This is the amount that is added. We have found that 0.0001 works well.

Be warned that the values of the ratio and the odds ratio will be affected by the amount specified here!

## Example1 - Finding Power

Two competing physical therapy treatments have been available for several years but have not yet been compared as to their effectiveness. The comparison of the two treatments is complicated by the sampling method that will be used. Instead of randomly assigning individuals to treatments, the researchers will randomly select two groups of physical therapists. The first group will be selected from those who use treatment 1. The second group will be selected from those who use treatment 2. The researchers will then follow up on the success or failure of the treatment for multiple patients of each physical therapist. The success rate of treatment 2 is known to be about 0.44. The researchers want to examine effect of the number of physical therapists used in each group and the number of patients for each physical therapist on the power of the test. They determine that the two treatments should be declared different if they are shown to be at least 0.07 apart in proportion. They plan to use the Farrington and Manning likelihood score test statistic to analyze the data. Based on similar studies, the intraclass correlation is estimated to be 0.02.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>0.0</b>
D1 .....	<b>0.07</b>
P2.....	<b>0.44</b>
K1.....	<b>10 15 20 25</b>
K2.....	<b>K1</b>
M1 .....	<b>10 to 50 by 10</b>
M2 .....	<b>M1</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
ICC.....	<b>0.02</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>
<b>Plot Setup Tab</b>	
Horizontal Axis Parameter .....	<b>K1</b>

## Annotated Output

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

### Numeric Results

#### Numeric Results of Tests Based on the Difference: P1 - P2

H0: P1-P2=D0. H1: P1-P2=D1<>D0. Test Statistic: Score test (Farrington & Manning)

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Alpha	Beta
Power	K1/M1	K2/M2	ICC							
0.1491	10/10	10/10	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.8509
0.2219	10/20	10/20	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.7781
0.2763	10/30	10/30	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.7237
0.3179	10/40	10/40	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.6821
0.3506	10/50	10/50	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.6494
0.2013	15/10	15/10	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.7987
0.3096	15/20	15/20	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.6904
0.3871	15/30	15/30	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.6129

.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

#### Summary Statements

Sample sizes of 100 in group one and 100 in group two, which were obtained by sampling 10 clusters with 10 subjects each in group one and 10 clusters with 10 subjects each in group two, achieve 15% power to detect a difference between the group proportions of 0.0700. The group two proportion is 0.4400. The group one proportion is assumed to be 0.4400 under the null hypothesis and 0.5100 under the alternative hypothesis. The test statistic used is the two-sided Score test (Farrington & Manning). The significance level of the test was 0.0500.

### Group 1 Clusters/Items: K1/M1

This line gives the value of K1, the number of clusters in group 1, followed by M1, the number of items per cluster in this group. The total number of items sampled in group 1 is  $N1 = K1 \times M1$ .

### Group 2 Clusters/Items: K2/M2

This line gives the value of K2, the number of clusters in group 2, followed by M2, the number of items per cluster in this group. The total number of items sampled in group 2 is  $N2 = K2 \times M2$ .

### Intraclass Corr.: ICC

This is the value of the intraclass correlation coefficient, ICC.

### Prop Grp 2 or Control: P2

This is the value of P2, the proportion responding positively in the control group.

### Prop|H0 Grp 1 or Trtmnt: P1.0

This is the value of P1.0, the proportion responding positively in the treatment group as specified by the null hypothesis. The difference between this value and P2 is the value used in the null hypothesis.

### Prop|H1 Grp 1 or Trtmnt: P1.1

This is the value of P1.1, the proportion responding positively in the treatment group as specified by the alternative hypothesis. The difference between this value and P2 is the value used in the alternative hypothesis.

## Diff if H0: D0

This is the value of  $D_0$ , the difference between proportions under the null hypothesis.

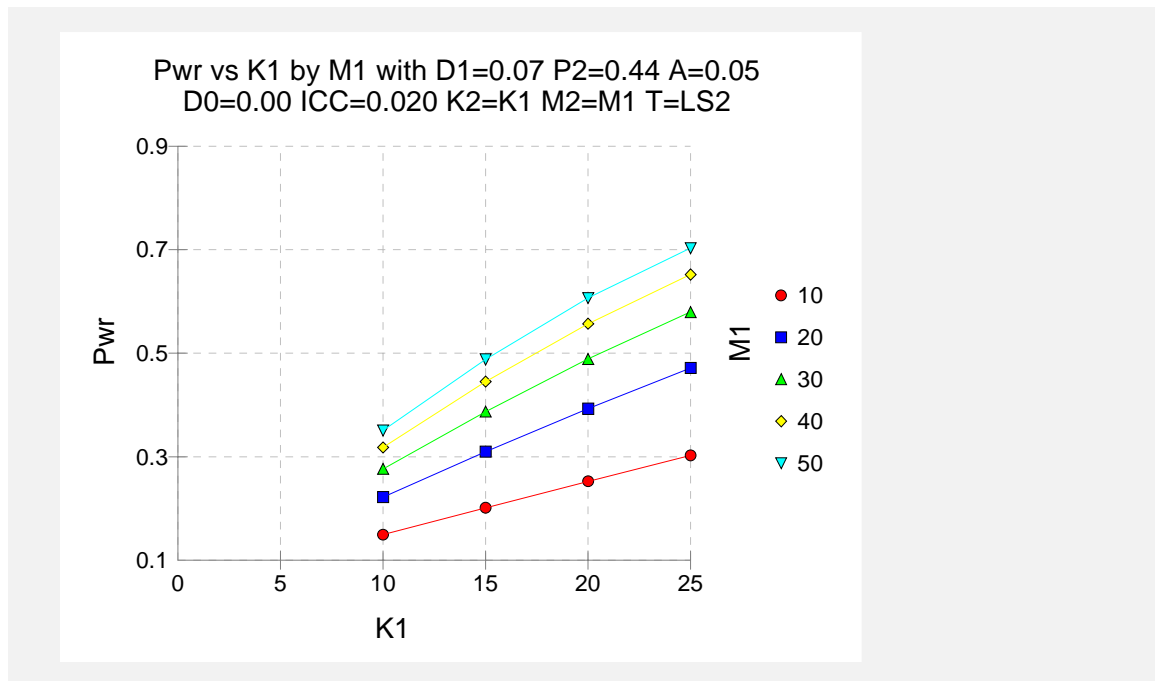
## Alpha

This is the value of alpha (significance level) that was targeted by the design.

## Beta

This is the value of beta, which is the probability of not rejecting a false null hypothesis.

## Plots Section



The values from the table are displayed on the above chart. This chart gives us a quick look at the power that is achieved for various combinations of cluster size and numbers of clusters.

## Example2 - Finding Sample Size

Continuing with Example1, the maximum number of therapists the researchers hope to use is 25 for each treatment. They decide to determine how many patients each therapist would have to treat to achieve 90% power if the maximum number of therapists is used.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>M1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>0.0</b>
D1.....	<b>0.07</b>
P2.....	<b>0.44</b>
K1 .....	<b>25</b>
K2.....	<b>K1</b>
M1 .....	<b>Ignored</b>
M2 .....	<b>M1</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
ICC .....	<b>0.02</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.10</b>

### Annotated Output

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

### Numeric Results

#### Numeric Results of Tests Based on the Difference: P1 - P2

H0: P1-P2=D0. H1: P1-P2=D1<>D0. Test Statistic: Score test (Farrington & Manning)

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 Control	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0	Diff if H1	Alpha	Beta
Power	K1/M1	K2/M2	ICC	P2			D0	D1		
0.9002	25/286	25/286	0.0200	0.4400	0.4400	0.5100	0.0000	0.0700	0.0500	0.0998

To achieve 90% power, each therapist would need to be evaluated on 286 patients.

## Example3 - Finding Power (Non-Inferiority)

A study is being designed to study the effectiveness of a new treatment. Historically, the standard treatment has enjoyed a 60% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the standard treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the standard treatment. The researchers will recommend adoption of the new treatment if it has a cure rate of at least 55%.

The researchers will recruit patients from various hospitals. All patients at a particular hospital will receive the same treatment. They anticipate an average of 100 patients per hospital.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data. They want to study the power of the one-sided Farrington and Manning test at group cluster sizes ranging from 2 to 10 for detecting a difference of -0.05 when the actual cure rate of the new treatment ranges from 60% to 66%. The significance level will be 0.05. Based on similar studies, they estimate the intracluster correlation to be 0.002.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>-0.05</b>
D1 .....	<b>0 .02 .04 .06</b>
P2.....	<b>0.6</b>
K1.....	<b>2 4 6 8 10</b>
K2.....	<b>K1</b>
M1 .....	<b>100</b>
M2 .....	<b>M1</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
ICC.....	<b>0.002</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>

## Annotated Output

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

### Numeric Results

#### Numeric Results of Tests Based on the Difference: $P_1 - P_2$

H0:  $P_1 - P_2 \leq D_0$ . H1:  $P_1 - P_2 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)

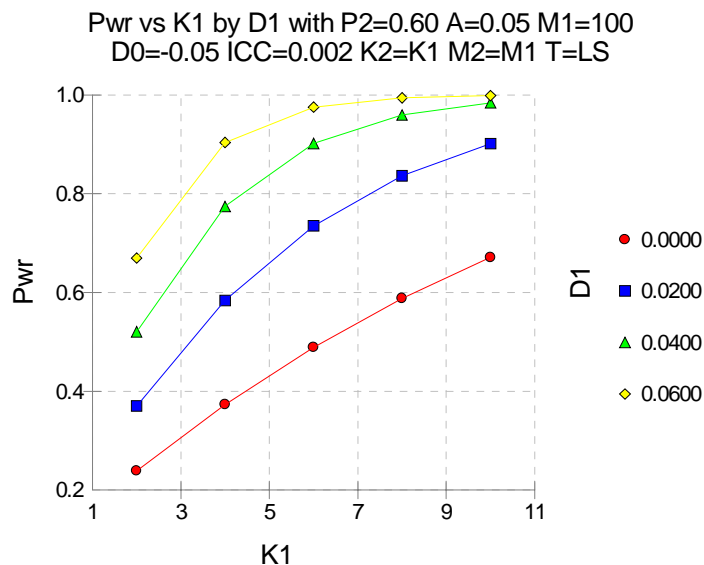
	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Alpha	Beta
Power	K1/M1	K2/M2	ICC							
0.2387	2/100	2/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.7613
0.3729	4/100	4/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.6271
0.4889	6/100	6/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.5111
0.5879	8/100	8/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.4121
0.6709	10/100	10/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.3291
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

#### Summary Statements

Sample sizes of 200 in group one and 200 in group two, which were obtained by sampling 2 clusters with 100 subjects each in group one and 2 clusters with 100 subjects each in group two, achieve 24% power to detect a difference between the group proportions of 0.0000. The group two proportion is 0.6000. The group one proportion is assumed to be 0.5500 under the null hypothesis and 0.6000 under the alternative hypothesis. The test statistic used is the one-sided Score test (Farrington & Manning). The significance level of the test was 0.0500.

This report shows the values of each of the parameters, one scenario per row. Most of the report columns have obvious interpretations. Those that may not be obvious are presented here.

### Plots Section



The values from the table are displayed on the above chart. This chart gives us a quick look at the sample size that will be required for various values of D1.

## Example4 - Finding the Sample Size (Non-Inferiority)

Continuing with the scenario given in Example 3, the researchers want to determine the number of clusters necessary for each value of D1 when the target power is set to 0.80.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. You can make these changes directly on your screen or you can load the template Example4 by clicking the Template tab and loading this template.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>K1</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>-0.05</b>
D1 .....	<b>0 .02 .04 .06</b>
P2 .....	<b>0.6</b>
K1 .....	<i>Ignored since this is the Find setting</i>
K2 .....	<b>K1</b>
M1 .....	<b>100</b>
M2 .....	<b>M1</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
ICC .....	<b>0.002</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 – P2**  
**H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 Control	Prop H0 Grp 1 or Trtmnt	Prop H1 Grp 1 or Trtmnt	Diff if H0	Diff if H1	Alpha	Beta
Power	K1/M1	K2/M2	ICC	P2	P1.0	P1.1	D0	D1		
0.8190	15/100	15/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.1810
0.8364	8/100	8/100	0.0020	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1636
0.8503	5/100	5/100	0.0020	0.6000	0.5500	0.6400	-0.0500	0.0400	0.0500	0.1497
0.8186	3/100	3/100	0.0020	0.6000	0.5500	0.6600	-0.0500	0.0600	0.0500	0.1814

The required sample size depends a great deal on the value of D1. The researchers should spend time determining the most appropriate value for D1.



## Example5 – Investigating the Impact of the Intraclass Correlation

Continuing with the scenario given in Example 4, the researchers, having decided that the most appropriate value of D1 is 0.02, now want to investigate the effect of the intraclass correlation on the sample size. From values found in other studies, they believe the ICC will be somewhere between 0.001 and 0.009.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>K1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.05</b>
D1.....	<b>0.02</b>
P2.....	<b>0.6</b>
K1 .....	<i>Ignored since this is the Find setting</i>
K2.....	<b>K1</b>
M1 .....	<b>100</b>
M2 .....	<b>M1</b>
Alternative Hypothesis .....	<b>One-Sided (H1:D1&gt;D0)</b>
ICC .....	<b>0.001 to 0.009 by 0.002</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

**Numeric Results of Tests Based on the Difference: P1 – P2**  
**H0: P1-P2<=D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Alpha	Beta
<b>Power</b>	<b>K1/M1</b>	<b>K2/M2</b>	<b>ICC</b>							
0.8207	7/100	7/100	0.0010	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1793
0.8099	8/100	8/100	0.0030	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1901
0.8020	9/100	9/100	0.0050	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1980
0.8275	11/100	11/100	0.0070	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1725
0.8197	12/100	12/100	0.0090	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1803

This chart shows that the necessary sample size almost doubles when the ICC is changed from 0.001 to 0.009. The researchers decide to obtain a narrower range for the value of ICC.

## Example6 - Validation using Donner and Klar

Donner (1997), page 63, present a sample size study in which  $P_2 = 0.06$ ,  $D_1 = -0.02$ ,  $D_0 = 0$ ,  $ICC = 0.01$ ,  $M_1 = M_2 = 100$ , two-sided  $\alpha = 0.05$ , and  $\beta = 0.20$ . Using the pooled z test statistic, they found the number of subjects to be 3698 in each group, which they round off to 38 clusters per group.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>K1</b>
Test Statistic .....	<b>Z test (pooled)</b>
D0 .....	<b>0.0</b>
D1 .....	<b>-0.02</b>
P2.....	<b>0.06</b>
K1.....	<i>Ignored since this is the Find setting</i>
K2.....	<b>K1</b>
M1 .....	<b>100</b>
M2 .....	<b>M1</b>
Alternative Hypothesis .....	<b>Two-Sided</b>
ICC.....	<b>0.01</b>
Alpha.....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

Numeric Results of Tests Based on the Difference:  $P_1 - P_2$   
 $H_0: P_1 - P_2 = D_0$ .  $H_1: P_1 - P_2 = D_1 \neq D_0$ . Test Statistic: Z test (pooled)

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 Control P2	Prop H0 Grp 1 or Trtmnt P1.0	Prop H1 Grp 1 or Trtmnt P1.1	Diff if H0 D0	Diff if H1 D1	Alpha	Beta
Power	K1/M1 38/100	K2/M2 38/100	ICC 0.0100	0.0600	0.0600	0.0400	0.0000	-0.0200	0.0500	0.1903

*PASS* has also found the required sample size to be 38 clusters.

## Chapter 235

# Non-Inferiority Tests of Two Proportions in Cluster Randomized Trials

## Introduction

This module provides power analysis and sample size calculation for non-inferiority and superiority tests in two-sample, cluster-randomized designs in which the outcome is binary.

## Three Procedures Documented Here

There are four procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, or ratios of proportions. Each of these options is listed separately on the menus.

## Technical Details

The methods contained in this module are identical to those discussed in the chapter “Comparing Proportions in Cluster Randomized Trials.” The input and output has simply been reformatted in a manner that is convenient for non-inferiority testing. A complete review of non-inferiority testing is given in the chapter “Non-Inferiority of Two Independent Proportions.” We refer you to these two chapters for complete technical details on the methods used in this module.

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers three procedures, each of which has different options. This section documents options that are common to all three procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *P1.1*, *Alpha*, *Beta*, *K1*, *M1*, or *ICC*. Under most situations, you will select either *Beta* or *K1*.

Select *K1* 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.

### Test Statistic

Specify which test statistic is used in searching and reporting. We recommend the likelihood score test.

### Higher Proportions Are

This option specifies whether proportions represent successes (better) or failures (worse).

#### Better (Successes)

When proportions represent successes, higher proportions are better. A noninferior treatment is one whose proportion is at least almost as high as that of the reference group.

For testing non-inferiority, D0 is negative, and R0 is less than one. For testing superiority, D0 is positive and R0 is greater than one.

#### Worse (Failures)

When proportions represent failures, lower proportions are better. A noninferior treatment is one whose proportion is at most almost as low as that of the reference group.

For testing non-inferiority, D0 is positive and R0 is greater than one. For testing superiority, D0 is negative, and R0 is less than one.

## P2 (Group 2 Proportion)

Specify the value of  $p_2$ , the control, baseline, or standard group's proportion. The null hypothesis is that the two proportions differ by a specified amount (See *Specify Group 1 Proportion using* below).

Since  $p_2$  is a proportion, these values must be between 0 and 1.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

## ICC (Intraclass Correlation)

Enter a value (or range of values) for the intraclass correlation. This correlation may be thought of as the simple correlation between any two observations in the same cluster. 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 between 0 to 1, typical values range from 0.002 to 0.05.

## K1 (Clusters Group 1)

Enter a value (or range of values) for the number of clusters in group one. You may enter a range of values such as *10 to 20 by 2*. The sample size for this group is equal to the number of clusters times the number of subjects per cluster.

## M1 (Items Group 1)

This is the average number of items (subjects) per cluster in group one. This value must be a positive number that is at least 1. You can use a list of values such as *100 150 200*.

## K2 (Clusters Group 2)

This is the number of clusters in group two. The sample size for this group is equal to the number of clusters times the number of subjects per cluster. This value must be a positive number.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *K1*, with no blanks. If you want to use *K1* directly, you do not have to pre-multiply by 1. For example, all of the following are valid entries: *10 K1 2K1 0.5K1*.

You can use a list of values such as *10 20 30* or *K1 2K1 3K1*.

## M2 (Items Group 2)

This is the number of items (subjects) per cluster in group two. This value must be a positive number.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *M1*, with no blanks. If you want to use *M1* directly, you do not have to pre-multiply by 1. For example, all of the following are valid entries: *10 M1 2M1 0.5M1*.

You can use a list of values such as *10 20 30* or *M1 2M1 3M1*.

## 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 unequal proportions when in fact they are not equal.

Values must be between 0 and 1. Because this is a one-sided test, unless you have good reason to do otherwise, the value of 0.025 is recommended. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.025 0.05* or *0.01 to 0.10 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 when in fact it is false.

Values must be between 0 and 1. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions, P1 and P2. P1.0 is the value of the P1 assumed by the null hypothesis and P1.1 is the value of P1 at which the power is calculated.

### P1.0 (Equiv. Proportion)

This option allows you to specify the value P1.0 directly. This is the value of the treatment group's proportion above which the treatment group is considered noninferior to the reference group. This option is only used for *Proportions*.

When *Higher Proportions Are* is set to *Better*, the trivial proportion is the smallest value of P1 for which the treatment group is declared noninferior to the reference group. In this case, P1.0 should be less than P2 for non-inferiority tests and greater than P2 for superiority tests. The reverse is the case when *Higher Proportions Are* is set to *Worse*.

Proportions must be between 0 and 1. They cannot take on the values 0 or 1. This value should not be set to exactly the value of P2.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

### P1.1 (Actual Proportion)

This option specifies the value of P1.1, the value of the treatment proportion at which the power is to be calculated. It is only used for *Proportions*. Proportions must be between 0 and 1. They cannot take on the values 0 or 1.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $D0$ , and  $D1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### D0 (Equiv.Difference)

This option specifies the trivial difference (often called the *margin of error*) between  $P1.0$  (the value of  $P1$  under  $H0$ ) and  $P2$ . This difference is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = P2 + D0$ . It is only used for *Differences*.

When *Higher Proportions Are* is set to *Better*, the trivial difference is that amount that  $P1$  can be less than  $P2$  and still have the treatment group declared noninferior to the reference group. In this case,  $D0$  should be negative for non-inferiority tests and positive for superiority tests.

The reverse is the case when *Higher Proportions Are* is set to *worse*.

You may enter a range of values such as  $-.03$   $-.05$   $-.10$  or  $-.05$  to  $-.01$  by  $.01$ . Differences must be between  $-1$  and  $1$ .  $D0$  cannot take on the values  $-1$ ,  $0$ , or  $1$ .

### D1 (Actual Difference)

This option specifies the actual difference between  $P1.1$  (the actual value of  $P1$ ) and  $P2$ . This is the value of the difference at which the power is calculated. In non-inferiority trials, this difference is often set to zero.

The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group). This difference is used with  $P2$  to calculate the value of  $P1.1$  using the formula:  $P1.1 = D1 + P2$ .

You may enter a range of values such as  $-.05$   $0$   $.5$  or  $-.05$  to  $.05$  by  $.02$ . Actual differences must be between  $-1$  and  $1$ . They cannot take on the values  $-1$  or  $1$ .

This option is only used for *Differences*.

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $R0$ , and  $R1$  are given, the values of  $P1.0$  and  $P1.1$  can be calculated.

### R0 (Equiv. Ratio)

This option specifies the trivial ratio (also called the Relative Margin of Equivalence) between  $P1.0$  and  $P2$ . The power calculations assume that  $P1.0$  is the value of  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0 = R0 \times P2$ . This option is only used for *Ratios*.

When *Higher Proportions Are* is set to *Better*, the trivial ratio is the relative amount by which  $P1$  can be less than  $P2$  and still have the treatment group declared noninferior to the reference group. In this case,  $R0$  should be less than 1 for non-inferiority tests and greater than 1 for superiority tests. The reverse is the case when 'Higher Proportions Are' is set to 'Worse'. In this case,  $R0$  should be less than 1 for non-inferiority tests and greater than 1 for superiority tests. The reverse is the case when *Higher Proportions Are* is set to *Worse*.

Ratios must be positive.  $R0$  cannot take on the value of 1.

You may enter a range of values such as *0.95 .97 .99* or *.91 to .99 by .02*.

### R1 (Actual Ratio)

This option specifies the ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the actual proportion in the treatment group. The power calculations assume that  $P1.1$  is the actual value of the proportion in group one. This difference is used with  $P2$  to calculate the value of  $P1.1$  using the formula:  $P1.1 = R1 \times P2$ . In non-inferiority trials, this ratio is often set to 1.

This option is only used for *Ratios*.

Ratios must be positive. You may enter a range of values such as *0.95 1 1.05* or *0.9 to 1.9 by 0.02*.



## Options Tab

The Options tab contains various limits and options.

### Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

### Zero Count Adjustment Method

Zero cell counts often cause calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option *Add to zero cells only*.

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts. We have found that adding 0.0001 seems to work well.

### Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This value indicates the amount that is added. We have found that 0.0001 works well.

Be warned that the value of the ratio and the odds ratio will be affected by the amount specified here!

## Example1 - Finding Power

A study is being designed to study the effectiveness of a new treatment. Historically, the standard treatment has enjoyed a 60% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the standard treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the standard treatment. The researchers will recommend adoption of the new treatment if it has a cure rate of at least 55%. That is, the margin of inferiority is -5%.

The researchers will recruit patients from various hospitals. All patients at a particular hospital will receive the same treatment. They anticipate an average of 100 patients per hospital. Based on similar studies, they estimate the intracluster correlation to be 0.002.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data. They want to study the power of the one-sided Farrington and Manning test at group cluster sizes ranging from 2 to 10 for detecting a difference of -0.05 when the actual cure rate of the new treatment ranges from 60% to 66%. The significance level will be 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0 .....	<b>-0.05</b>
D1 .....	<b>0 .02 .04 .06</b>
P2.....	<b>0.6</b>
K1.....	<b>2 4 6 8 10</b>
M1 .....	<b>100</b>
K2.....	<b>K1</b>
M2 .....	<b>M1</b>
Higher Proportions Are .....	<b>Better</b>
ICC.....	<b>0.002</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>

## Annotated Output

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

### Numeric Results

**Numeric Results for Non-Inferiority Tests Based on the Difference: P1 - P2**  
**H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Grp 2 Prop P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Alpha	Beta
Power	K1/M1	K2/M2	ICC							
0.2387	2/100	2/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.7613
0.3729	4/100	4/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.6271
0.4889	6/100	6/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.5111
0.5879	8/100	8/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.4121
0.6709	10/100	10/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.3291
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

#### Summary Statements

Sample sizes of 200 in group one and 200 in group two, which were obtained by sampling 2 clusters with 100 subjects each in group one and 2 clusters with 100 subjects each in group two, achieve 24% power to detect a non-inferiority margin difference between the group proportions of -0.0500. The group two proportion is 0.6000. The group one proportion is assumed to be 0.5500 under the null hypothesis and 0.6000 under the alternative hypothesis. The test statistic used is the one-sided Score test (Farrington & Manning). The significance level of the test was 0.0500.

This report shows the values of each of the parameters, one scenario per row. Most of the report columns have obvious interpretations. Those that may not be obvious are presented here.

### Group 1 Clusters/Items: K1/M1

This line gives the value of K1, the number of clusters in group 1, followed by M1, the number of items per cluster in this group. The total number of items sampled in group 1 is  $N1 = K1 \times M1$ .

### Group 2 Clusters/Items: K2/M2

This line gives the value of K2, the number of clusters in group 2, followed by M2, the number of items per cluster in this group. The total number of items sampled in group 2 is  $N2 = K2 \times M2$ .

### Intraclass Corr.: ICC

This is the value of the intraclass correlation coefficient, ICC.

### Prop Grp 2: P2

This is the value of P2, the proportion responding positively in the control group.

### Equiv. Grp 1 Prop P1.0

This is the value of P1.0, the response rate of the treatment group, as specified by the null hypothesis of inferiority. Values of P1 less than this amount are considered different from P2. Values of P1 greater than this are considered noninferior to the reference group. The difference between this value and P2 is the value of the null hypothesis.

## Actual Grp 1 Prop P1.1

This is the value of P1.1, the response rate of the treatment group, at which the power is computed. This is the value of P1 under the alternative hypothesis. The difference between this value and P2 is the value of the alternative hypothesis.

## Equiv. Margin Diff D0

This is the value of D0, the difference between the two group proportions under the null hypothesis. This value is often called the *margin of non-inferiority*.

## Actual Margin Diff D1

This is the value of D1, the difference between the two group proportions at which the power is computed. This is the value of the difference under the alternative hypothesis.

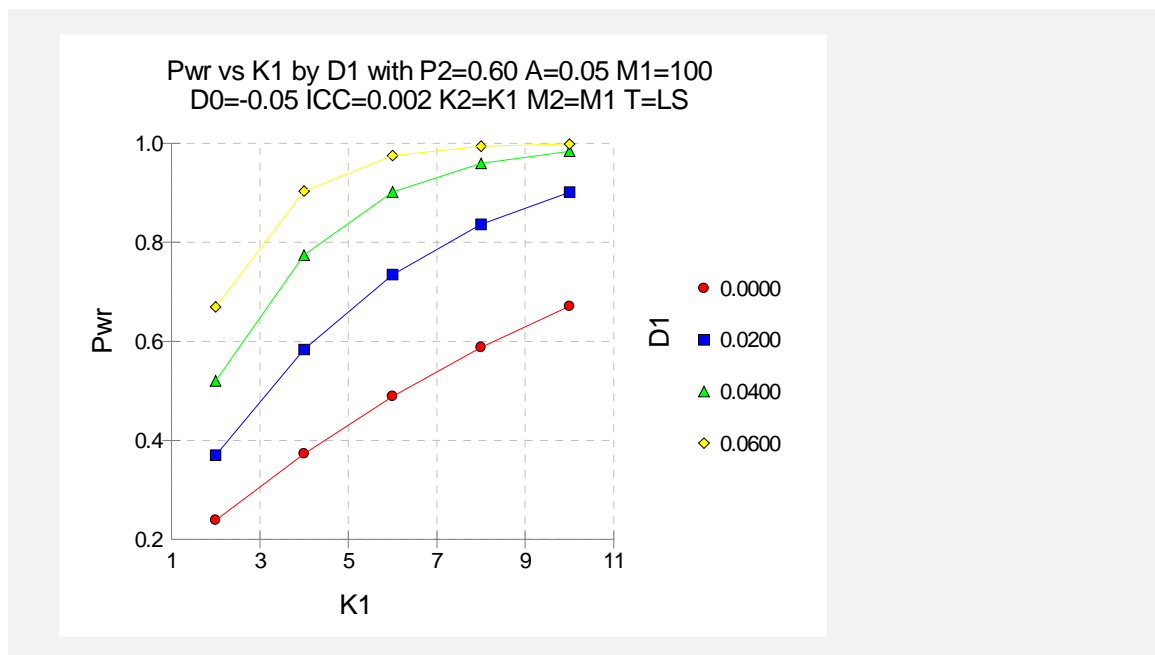
## Alpha

This is the value of alpha (significance level) that was targeted by the design.

## Beta

This is the value of beta, which is the probability of not rejecting a false null hypothesis.

## Plots Section



The values from the table are displayed on the above chart. This chart gives us a quick look at the sample sizes that will be required for various values of D1.

## Example2 - Finding the Sample Size (Number of Clusters)

Continuing with the scenario given in Example 1, the researchers want to determine the number of clusters necessary for each value of D1 when the target power is set to 0.80.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>K1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.05</b>
D1.....	<b>0 .02 .04 .06</b>
K1 .....	<i>Ignored since this is the Find setting</i>
M1 .....	<b>100</b>
K2.....	<b>K1</b>
M2 .....	<b>M1</b>
Higher Proportions Are.....	<b>Better</b>
P2.....	<b>0.6</b>
ICC .....	<b>0.002</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

**Numeric Results for Non-Inferiority Tests Based on the Difference: P1 - P2**  
**H0: P1-P2≤D0. H1: P1-P2=D1>D0. Test Statistic: Score test (Farrington & Manning)**

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr. ICC	Grp 2 Prop P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Alpha	Beta
<b>Power</b>	<b>K1/M1</b>	<b>K2/M2</b>								
0.8190	15/100	15/100	0.0020	0.6000	0.5500	0.6000	-0.0500	0.0000	0.0500	0.1810
0.8364	8/100	8/100	0.0020	0.6000	0.5500	0.6200	-0.0500	0.0200	0.0500	0.1636
0.8503	5/100	5/100	0.0020	0.6000	0.5500	0.6400	-0.0500	0.0400	0.0500	0.1497
0.8186	3/100	3/100	0.0020	0.6000	0.5500	0.6600	-0.0500	0.0600	0.0500	0.1814

The required sample size depends a great deal on the value of D1. The researchers should spend time determining the most appropriate value for D1.

## Example3 - Validation

We could not find an example of this type of analysis in the literature. Therefore, we will validate the procedure by comparing the results to those given in Example3 in the chapter “Comparing Proportions in Cluster Randomized Trials,” since both modules should give identical results. Validation can be accomplished by running Example 1 in this chapter and Example3 in that chapter. If you do this, you will see that both procedures give the same results.

## Example4 - Finding Power after an Experiment

A group of researchers want to show that a new, less expensive treatment works at least as well as the current treatment. They believe, in fact, that the new treatment is about 0.10 higher in proportion of success. One hundred patients at each of 10 randomly chosen hospitals were given the current treatment. One hundred patients at each of 10 randomly chosen hospitals were given the new treatment. It was agreed before the experiment that the new treatment needed to be no less than 0.05 in proportion of success below the current treatment to be considered noninferior. The proportion of patients responding to the current treatment was  $821/1000 = 0.821$ . The proportion of patients responding to the new treatment was  $819/1000 = 0.819$ . This result did not show significant noninferiority at the 0.05 level. The researchers want to know the power of their noninferiority test. They decide to use the intracluster correlation coefficient estimated from the data, which was 0.0068. Although the observed difference in proportions is  $0.819 - 0.821 = -0.002$ , the trivial difference is still -0.05. This value is used in the power calculation.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.....	<b>-0.05</b>
D1.....	<b>0.0 0.10</b>
P2.....	<b>0.821</b>
K1.....	<b>10</b>
K2.....	<b>K1</b>
M1 .....	<b>100</b>
M2 .....	<b>M1</b>
Higher Proportions Are.....	<b>Better</b>
ICC .....	<b>0.0068</b>
Alpha .....	<b>0.05</b>
Beta.....	<i>Ignored</i>

### Annotated Output

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

## Numeric Results

Numeric Results for Non-Inferiority Tests Based on the Difference:  $P_1 - P_2$   
 $H_0: P_1 - P_2 \leq D_0$ .  $H_1: P_1 - P_2 = D_1 > D_0$ . Test Statistic: Score test (Farrington & Manning)

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Grp 2 Prop P2	Equiv. Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Equiv. Margin Diff D0	Actual Margin Diff D1	Alpha	Beta
Power	K1/M1	K2/M2	ICC							
0.7272	10/100	10/100	0.0068	0.8210	0.7710	0.8210	-0.0500	0.0000	0.0500	0.2728
1.0000	10/100	10/100	0.0068	0.8210	0.7710	0.9210	-0.0500	0.1000	0.0500	0.0000

If indeed the new treatment were 0.10 higher in proportion of success, the power for showing noninferiority would be 1.0000. If the true proportions are the same, the power would be 0.7272.



## Example5 - Finding Sample Size (Individuals Within Clusters)

An agency would like to show the proportion of success of a new treatment is no less than that of the current treatment. Thirty doctors are available for the study. Fifteen will be randomly chosen to be trained to administer the new treatment. The remaining fifteen will continue to administer the current treatment. The new treatment will be considered noninferior if the the proportion of success is at least 90% of the current treatment success. The agency would like to know the number of patients that need to be treated by each doctor to achieve 80% power for the noninferiority test. Various values for the intraclass correlation coefficient will be use since its true value is unknown. It is expected that the two treatments will have a success rate near 0.65. Alpha is set at 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>M1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
R0 (Trivial Ratio) .....	<b>0.90</b>
R1 (Actual Ratio) .....	<b>1.0</b>
P2.....	<b>0.65</b>
K1 .....	<b>15</b>
K2.....	<b>K1</b>
M1 .....	<i>Ignored since this is the find setting</i>
M2 .....	<b>M1</b>
Higher Proportions Are.....	<b>Better</b>
ICC .....	<b>0.001 to 0.01 by 0.001</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

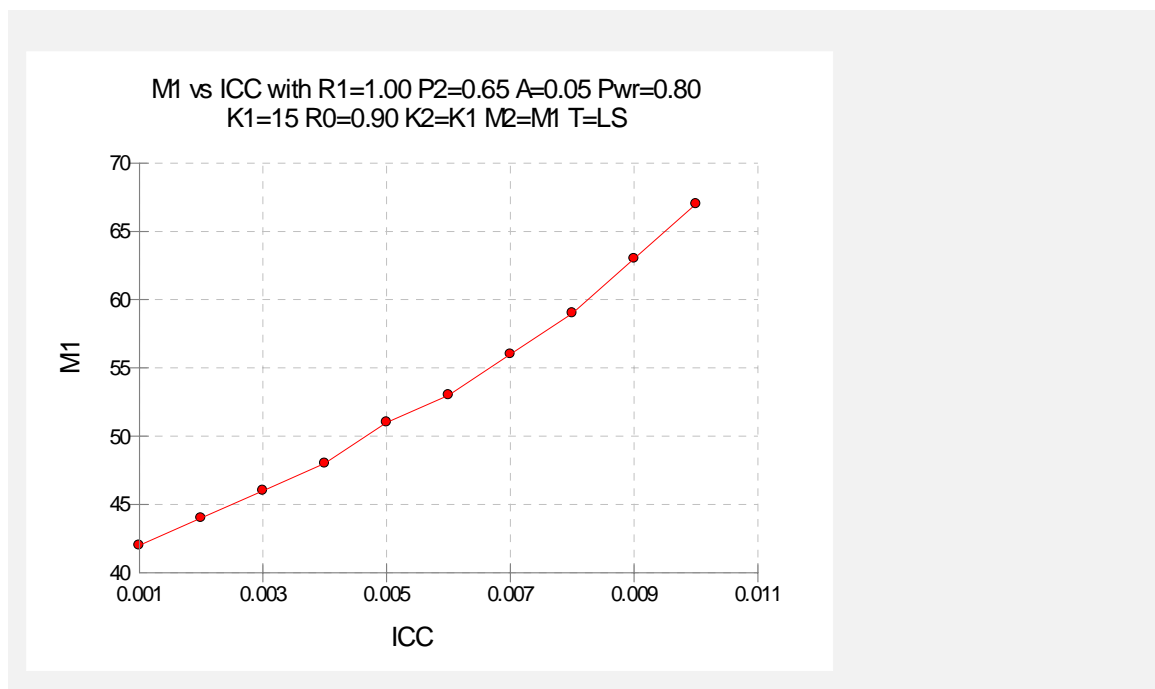
Click the Run button to perform the calculations and generate the following output.

## Numeric Results

Numeric Results for Non-Inferiority Tests Based on the Ratio:  $P1 / P2$   
 $H0: P1/P2 \leq R0$ .  $H1: P1/P2 = R1 > R0$ . Test Statistic: Score test (Farrington & Manning)

Power	Group 1 Clusters/ Items K1/M1	Group 2 Clusters/ Items K2/M2	Intra- Cluster Corr. ICC	Grp 2 Prop P2	Trivial Grp 1 Prop P1.0	Actual Grp 1 Prop P1.1	Trivial Margin Ratio R0	Actual Margin Ratio R1	Alpha	Beta
0.8011	15/42	15/42	0.0010	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1989
0.8023	15/44	15/44	0.0020	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1977
0.8023	15/46	15/46	0.0030	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1977
0.8017	15/48	15/48	0.0040	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1983
0.8045	15/51	15/51	0.0050	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1955
0.8011	15/53	15/53	0.0060	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1989
0.8017	15/56	15/56	0.0070	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1983
0.8005	15/59	15/59	0.0080	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1995
0.8017	15/63	15/63	0.0090	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1983
0.8011	15/67	15/67	0.0100	0.6500	0.5850	0.6500	0.900	1.000	0.0500	0.1989

## Plots Section



The number of patients that should be seen by each doctor ranges from 42 to 67, depending on the intraclass correlation coefficient..

## Chapter 240

# Equivalence of Two Proportions in Cluster-Randomized Trials

## Introduction

This module provides power analysis and sample size calculation for equivalence tests in two-sample, cluster-randomized designs in which the outcome is binary.

## Three Procedures Documented Here

There are three procedures in the menus that use the program module described in this chapter. These procedures are identical except for the type of parameterization. The parameterization can be in terms of proportions, differences in proportions, or ratios of proportions. Each of these options is listed separately on the menus.

## Technical Details

The methods contained in this module are identical to those discussed in the chapter entitled “Comparing Proportions in Cluster-Randomized Trials.” The input and output has simply been reformatted in a manner that is convenient for equivalence testing. A complete review of equivalence testing is given in the chapter “Equivalence of Two Independent Proportions.” We refer you to these two chapters for complete technical details on the methods used in this module.

## 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 (Common Options)

The Data tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and beta. This chapter covers three procedures, each of which has different options. This section documents options that are common to all three procedures. Later, unique options for each procedure will be documented.

### Find

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *P1.1*, *Alpha*, *Beta*, *K1*, *M1*, or *ICC*. Under most situations, you will select either *Beta* or *K1*.

Select *K1* 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.

### Test Statistic

Specify which test statistic is used in searching and reporting. We recommend the likelihood score test.

### P2 (Control Group Proportion)

Specify the value of P2, the control, baseline, or standard group's proportion. The null hypothesis is that the two proportions differ by a specified amount (See *Specify Group 1 Proportion using below*).

Since P2 is a proportion, these values must be between 0 and 1.

You may enter a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

### ICC (Intraclass Correlation)

Enter a value (or range of values) for the intraclass correlation. This correlation may be thought of as the simple correlation between any two observations in the same cluster. 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 0 to 1, typical values range from 0.002 to 0.05.

### K1 (Clusters Group 1)

Enter a value (or range of values) for the number of clusters in this group. You may enter a range of values such as *10 to 20 by 2*. The sample size for this group is equal to the number of clusters times the number of subjects per cluster.

## M1 (Items Group 1)

This is the average number of items (subjects) per cluster in group one. This value must be a positive number that is at least 1. You can use a list of values such as *100 150 200*.

## K2 (Clusters Group 2)

This is the number of clusters in group two. The sample size for this group is equal to the number of clusters times the number of subjects per cluster. This value must be a positive number.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *K1*, with no blanks. If you want to use *K1* directly, you do not have to pre-multiply by 1. For example, all of the following are valid entries: *10 K1 2K1 0.5K1*.

You can use a list of values such as *10 20 30* or *K1 2K1 3K1*.

## M2 (Items Group 2)

This is the number of items (subjects) per cluster in group two. This value must be a positive number.

If you simply want a multiple of the value for group one, you would enter the multiple followed by *M1*, with no blanks. If you want to use *M1* directly, you do not have to pre-multiply by 1. For example, all of the following are valid entries: *10 M1 2M1 0.5M1*.

You can use a list of values such as *10 20 30* or *M1 2M1 3M1*.

## 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 unequal proportions when in fact they are not equal.

Values must be between 0 and 1. Because this is a two-sided test, unless you have good reason to do otherwise, the value of 0.05 is recommended. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.025 0.05* or *0.01 to 0.10 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 when in fact it is false.

Values must be between 0 and 1. Historically, the value of 0.20 was often 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 1-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 power values of 0.95, 0.90, and 0.80, respectively.

## Data Tab (Proportion)

This section documents options that are used when the parameterization is in terms of the values of the two proportions,  $P_1$  and  $P_2$ .  $P_{1.0}$  is the value of the  $P_1$  assumed by the null hypothesis and  $P_{1.1}$  is the value of  $P_1$  at which the power is calculated.

### P1.0U & P1.0L (Upper & Lower Equivalence Proportion)

Specify the *margin of equivalence* directly by giving the upper and lower bounds of  $P_{1.0}$ . The two groups are assumed to be equivalent when  $P_{1.0}$  is between these values. Thus,  $P_{1.0U}$  should be greater than  $P_2$  and  $P_{1.0L}$  should be less than  $P_2$ .

This option is only used for *Proportions*.

Note that the values of  $P_{1.0U}$  and  $P_{1.0L}$  are used in pairs. Thus, the first values of  $P_{1.0U}$  and  $P_{1.0L}$  are used together, and then the second values of each are used, and so on.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

Proportions must be between 0 and 1. They cannot take on the values 0 or 1. These values should surround  $P_2$ .

### P1.1 (Actual Proportion)

This option specifies the value of  $P_{1.1}$ , which is the value of the treatment proportion at which the power is to be calculated. It is only used for *Proportions*. Proportions must be between 0 and 1. They cannot take on the values 0 or 1.

You may enter a range of values such as *0.03 0.05 0.10* or *0.01 to 0.05 by 0.01*.

## Data Tab (Difference)

This section documents options that are used when the parameterization is in terms of the difference,  $P1 - P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $D0$ , and  $D1$  are given, the values of  $P1.1$  and  $P1.0$  can be calculated.

### D0.U & D0.L (Upper & Lower Equivalence Difference)

Specify the *margin of equivalence* by specifying the largest distance above ( $D0.U$ ) and below ( $D0.L$ )  $P2$  which will still result in the conclusion of equivalence. As long as the actual difference is between these two values, the difference is not large enough to be of practical importance.

The values of  $D0.U$  must be positive and the values of  $D0.L$  must be negative.  $D0.L$  can be set to  $-'D0.U'$ , which is usually what is desired.

The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0U = D0.U + P2$ .

This option is only used for *Differences*.

You may enter a range of values for  $D0.U$  such as  $.03 .05 .10$  or  $.05 to .20 by .05$ .

Note that if you enter values for  $D0.L$  (other than  $-'D0.U'$ ), they are used in pairs with the values of  $D0.U$ . Thus, the first values of  $D0.U$  and  $D0.L$  are used together, then the second values of each are used, and so on.

RANGE:

$D0.L$  must be between  $-1$  and  $0$ .  $D0.U$  must be between  $0$  and  $1$ . Neither can take on the values  $-1$ ,  $0$ , or  $1$ .

### D1 (Actual Difference)

This option specifies the actual difference between  $P1.1$  (the actual value of  $P1$ ) and  $P2$ . This is the value of the difference at which the power is calculated. In equivalence trials, this difference is often set to zero.

The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1 (experimental or treatment group). This difference is used with  $P2$  to calculate the true value of  $P1$  using the formula:  $P1.1 = D1 + P2$ .

You may enter a range of values such as  $-.05 0 .5$  or  $-.05 to .05 by .02$ . Actual differences must be between  $-1$  and  $1$ . They cannot take on the values  $-1$  or  $1$ .

This option is only used for *Differences*.

## Data Tab (Ratio)

This section documents options that are used when the parameterization is in terms of the ratio,  $P1 / P2$ .  $P1.0$  is the value of  $P1$  assumed by the null hypothesis and  $P1.1$  is the value of  $P1$  at which the power is calculated. Once  $P2$ ,  $R0$ , and  $R1$  are given, the values of  $P1.0$  and  $P1.1$  can be calculated.

### R0.U & R0.L (Upper & Lower Equivalence Ratio)

Specify the *margin of equivalence* by specifying the largest ratio ( $P1/P2$ ) above ( $R0.U$ ), and below ( $R0.L$ ), which will still result in the conclusion of equivalence. As long as the actual ratio is between these two values, the difference between the proportions is not said to be large enough to be of practical importance.

The values of  $R0.U$  must be greater than 1 and the values of  $R0.L$  must be less than 1.  $R0.L$  can be set to ' $1/R0.U$ ', which is most often desired.

The power calculations assume that  $P1.0$  is the value of the  $P1$  under the null hypothesis. This value is used with  $P2$  to calculate the value of  $P1.0$  using the formula:  $P1.0U = R0.U \times P2$ .

This option is only used for *Ratios*.

You may enter a range of values for  $R0.U$  such as *1.1 1.5 1.8* or *1.1 to 2.1 by 0.2*.

Note that if you enter values for  $R0.L$  (other than ' $1/R0.U$ '), they are used in pairs with the values of  $R0.U$ . Thus, the first values of  $R0.U$  and  $R0.L$  are used together, then the second values of each are used, and so on.

$R0.L$  must be between 0 and 1.  $R0.U$  must be greater than 1. Neither can take on the value 1.

### R1 (Actual Ratio)

This option specifies the ratio of  $P1.1$  and  $P2$ , where  $P1.1$  is the actual proportion in the treatment group. The power calculations assume that  $P1.1$  is the actual value of the proportion in group 1. This difference is used with  $P2$  to calculate the value of  $P1$  using the formula:  $P1.1 = R1 \times P2$ . In equivalence trials, this ratio is often set to 1.

This option is only used for *Ratios*.

Ratios must be positive. You may enter a range of values such as *0.95 1 1.05* or *0.9 to 1.9 by 0.02*.



# Options Tab

The Options tab contains various limits and options.

## Maximum Iterations

Specify the maximum number of iterations before the search for the criterion of interest is aborted. When the maximum number of iterations is reached without convergence, the criterion is not reported. A value of at least 500 is recommended.

## Zero Count Adjustment Method

Zero cell counts often cause calculation problems. To compensate for this, a small value (called the Zero Count Adjustment Value) can be added either to all cells or to all cells with zero counts. This option specifies whether you want to use the adjustment and which type of adjustment you want to use. We recommend that you use the option *Add to zero cells only*.

Zero cell values often do not occur in practice. However, since power calculations are based on total enumeration, they will occur in power and sample size estimation.

Adding a small value is controversial, but can be necessary for computational considerations. Statisticians have recommended adding various fractions to zero counts. We have found that adding 0.0001 seems to work well.

## Zero Count Adjustment Value

Zero cell counts cause many calculation problems when computing power or sample size. To compensate for this, a small value may be added either to all cells or to all zero cells. This value indicates the amount that is added. We have found that 0.0001 works well.

Be warned that the value of the ratio and the odds ratio will be affected by the amount specified here!

## Example1 - Finding Power

A study is being designed to establish the equivalence of a new treatment compared to the current treatment. Historically, the standard treatment has enjoyed a 60% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the standard treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the standard treatment. The researchers will recommend adoption of the new treatment if its cure rate is within 0.15 of the standard treatment.

The researchers will recruit patients from various hospitals. All patients at a particular hospital will receive the same treatment. They anticipate enlisting an average of 50 patients per hospital. Based on similar studies, they estimate the intracluster correlation to be 0.002.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data. They want to study the power of the two, one-sided tests proposed by Farrington and Manning when the number of clusters per groups ranges from 2 to 10. They want to investigate the behavior of this test when the actual cure rate of the new treatment ranges from 60% to 66%. The significance level will be 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic .....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.15</b>
D0.L .....	<b>-D0.U</b>
D1 .....	<b>0 .03 .06</b>
P2.....	<b>0.6</b>
K1.....	<b>2 4 6 8 10</b>
M1 .....	<b>50</b>
K2.....	<b>K1</b>
M2 .....	<b>M1</b>
ICC .....	<b>0.002</b>
Alpha.....	<b>0.05</b>
Beta.....	<i>Ignored since this is the Find setting</i>

## Annotated Output

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

### Numeric Results

**Numeric Results for Equivalence Tests Based on the Difference: P1 - P2**

**H0: P1-P2<=D0L or P1-P2>=D0U. H1: D0L<P1-P2=D1<D0U.**

**Test Statistic: Score tests**

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Alpha	Beta
<b>Power</b>	<b>K1/M1</b>	<b>K2/M2</b>	<b>ICC</b>								
0.3459	2/50	2/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0000	0.0500	0.6541
0.8065	4/50	4/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0000	0.0500	0.1935
0.9494	6/50	6/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0000	0.0500	0.0506
0.9879	8/50	8/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0000	0.0500	0.0121
0.9973	10/50	10/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0000	0.0500	0.0027
0.3279	2/50	2/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0300	0.0500	0.6721
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

#### Summary Statements

Sample sizes of 100 in group one and 100 in group two, which were obtained by sampling 2 clusters with 50 subjects each in group one and 2 clusters with 50 subjects each in group two, achieve 33% power to detect equivalence. The margin of equivalence, given in terms of the difference between the proportions, extends from -0.1500 to 0.1500. The actual difference between the proportions is 0.0000. The group two proportion is 0.6000. The calculations assume that two, one-sided z tests (unpooled) were used. The significance level of the test was 0.0500.

This report shows the values of each of the parameters, one scenario per row. Most of the report columns have obvious interpretations. Those that may not be obvious are presented here.

### Group 1 Clusters/Items: K1/M1

This line gives the value of K1, the number of clusters in group 1, followed by M1, the number of items per cluster in this group. The total number of items sampled in group 1 is  $N1 = K1 \times M1$ .

### Group 2 Clusters/Items: K2/M2

This line gives the value of K2, the number of clusters in group 2, followed by M2, the number of items per cluster in this group. The total number of items sampled in group 2 is  $N2 = K2 \times M2$ .

### Intraclass Corr.: ICC

This is the value of the intraclass correlation coefficient, ICC.

### Prop Grp 2: P2

This is the value of P2, the proportion responding positively in the control group.

### Lower & Upper Equiv. Grp 1 Prop: P1.0L & P1.0U

These are the margin of equivalence for the response rate of the treatment group, as specified by the null hypothesis of non-equivalence. Values of P1 inside these limits are considered equivalent to P2.

## Lower & Upper Equiv. Margin Diff: D0.L & D0.U

These set the margin of equivalence for the difference in response rates. Values of the difference outside these limits are considered *non-equivalent*.

## Actual Margin Diff D1

This is the value of D1, the difference between the two group proportions at which the power is computed. This is the value of the difference under the alternative hypothesis.

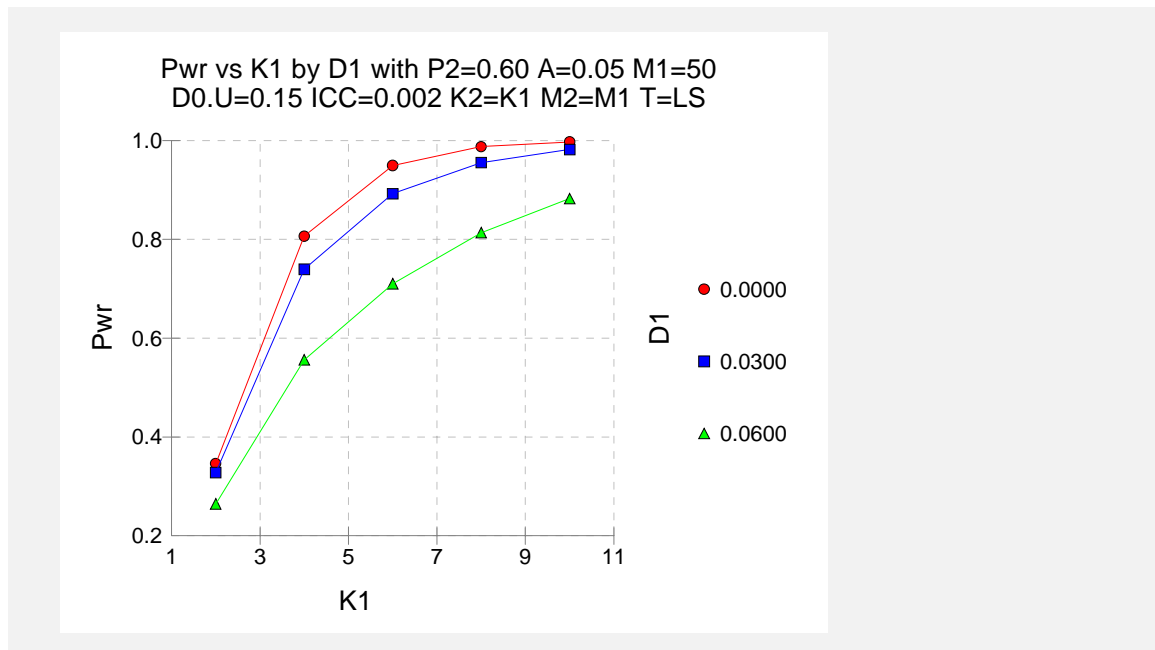
## Alpha

This is the value of alpha (significance level) that was targeted by the design.

## Beta

This is the value of beta, which is the probability of not rejecting a false null hypothesis.

## Plots Section



The values from the table are displayed on the above chart. This chart gives us a quick look at the sample size that will be required for various values of D1.

## Example2 - Finding the Sample Size (Number of Clusters)

Continuing with the scenario given in Example 1, the researchers want to determine the number of clusters necessary for each value of D1 when the target power is set to 0.80.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>K1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.15</b>
D0.L.....	<b>-D0.U</b>
D1.....	<b>0 .03 .06</b>
P2.....	<b>0.6</b>
K1 .....	<i>Ignored since this is the Find setting</i>
M1 .....	<b>50</b>
K2.....	<b>K1</b>
M2 .....	<b>M1</b>
ICC .....	<b>0.002</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

Numeric Results for Equivalence Tests Based on the Difference: P1 - P2  
H0:  $P1 - P2 \leq D0L$  or  $P1 - P2 \geq D0U$ . H1:  $D0L < P1 - P2 = D1 < D0U$ .  
Test Statistic: Score tests

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2	Lower Equiv. Grp 1 Prop	Upper Equiv. Grp 1 Prop	Lower Equiv. Margin Diff	Upper Equiv. Margin Diff	Actual Margin Diff		
Power	K1/M1	K2/M2	ICC	P2	P1.0L	P1.0U	D0.L	D0.U	D1	Alpha	Beta
0.8065	4/50	4/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0000	0.0500	0.1935
0.8324	5/50	5/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0300	0.0500	0.1676
0.8137	8/50	8/50	0.0020	0.6000	0.4500	0.7500	-0.1500	0.1500	0.0600	0.0500	0.1863

The required sample size depends a great deal on the value of D1. The researchers should spend time determining the most appropriate value for D1.

## Example3 - Validation

We could not find an example of this type of analysis in the literature. Therefore, we will validate the procedure by comparing the results to those given in the chapter entitled “Equivalence – Two Independent Proportions,” since both modules should give identical results for the same sample sizes when the ICC is set to zero. We ran the case when  $N1 = N2 = 200$ ,  $P2 = 0.6$ ,  $D0.U = 0.15$ ,  $D1 = 0$ , and  $\text{Alpha} = 0.05$ . In this module, set  $M1 = 1$  and set  $K1 = 200$ . Both program modules calculated the power to be 0.8482 in this case.

## Example4 - Finding Power after an Experiment

Individuals promoting a new, more expensive treatment claim that it achieves better results than the current treatment without citing statistical evidence. A group of researchers attempted to show the claim was false through a study involving 12 hospitals. Two hundred patients at each of 6 randomly chosen hospitals were given the current treatment. Two hundred patients at each of the remaining 6 hospitals were given the new treatment. It was agreed before the experiment that if a difference of less than 0.05 in proportion of success could be shown, the two treatments would be deemed equivalent. The proportion of patients responding properly to the current treatment was  $540/1200 = 0.450$ . The proportion of patients responding properly to the new treatment was  $570/1200 = 0.475$ . This result did not show significant equivalence at the 0.05 level. The researchers want to know the power of their equivalence test. They decide to use the intracluster correlation coefficient estimated from the data, which was 0.0043. Although the observed difference in proportions is  $0.475 - 0.450 = 0.025$ , the equivalence difference is still 0.05. This value is used in the power calculation.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.05</b>
D0.L.....	<b>-D0.U</b>
D1 .....	<b>0.0</b>
P2 .....	<b>0.45</b>
K1 .....	<b>6</b>
K2 .....	<b>K1</b>
M1 .....	<b>200</b>
M2 .....	<b>M1</b>
ICC .....	<b>0.0043</b>
Alpha .....	<b>0.05</b>
Beta .....	<i>Ignored since this is the Find setting</i>

## Annotated Output

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

### Numeric Results

Numeric Results for Equivalence Tests Based on the Difference: P1 - P2

H0:  $P1 - P2 \leq D0.L$  or  $P1 - P2 \geq D0.U$ . H1:  $D0.L < P1 - P2 = D1 < D0.U$ .

Test Statistic: Score tests

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr.	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Alpha	Beta
Power	K1/M1	K2/M2	ICC								
0.1309	6/200	6/200	0.0043	0.4500	0.4000	0.5000	-0.0500	0.0500	0.0000	0.0500	0.8691

The power of the test of equivalence was only 0.1309.



## Example5 - Finding Sample Size (Individuals within Clusters)

An agency would like to show the proportion of success is the same for two treatments. Eight doctors are available for the study. Four will be randomly chosen to be trained to administer treatment 1. The remaining four will administer treatment 2. The treatments will be considered equivalent if the the proportion of success of treatment 1 is within 0.10 of treatment 2 success. The agency would like to know the number of patients that need to be treated by each doctor to achieve 80% power for the equivalence test. Various values for the intracluster correlation coefficient will be use since its true value is unknown. It is expected that the two treatments will have a success rate near 0.70. Alpha is set at 0.05.

### Setup

This section presents the values of each of the parameters needed to run this example. First, load the procedure. This example uses the difference parameterization. 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>M1</b>
Test Statistic.....	<b>Likelihood Score (Farr. &amp; Mann.)</b>
D0.U .....	<b>0.10</b>
D0.L.....	<b>-D0.U</b>
D1.....	<b>0.0</b>
P2.....	<b>0.70</b>
K1.....	<b>4</b>
K2.....	<b>K1</b>
M1 .....	<i>Ignored since this is the find setting</i>
M2 .....	<b>M1</b>
ICC .....	<b>0.001 to 0.01 by 0.001</b>
Alpha .....	<b>0.05</b>
Beta.....	<b>0.20</b>

## Annotated Output

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

### Numeric Results

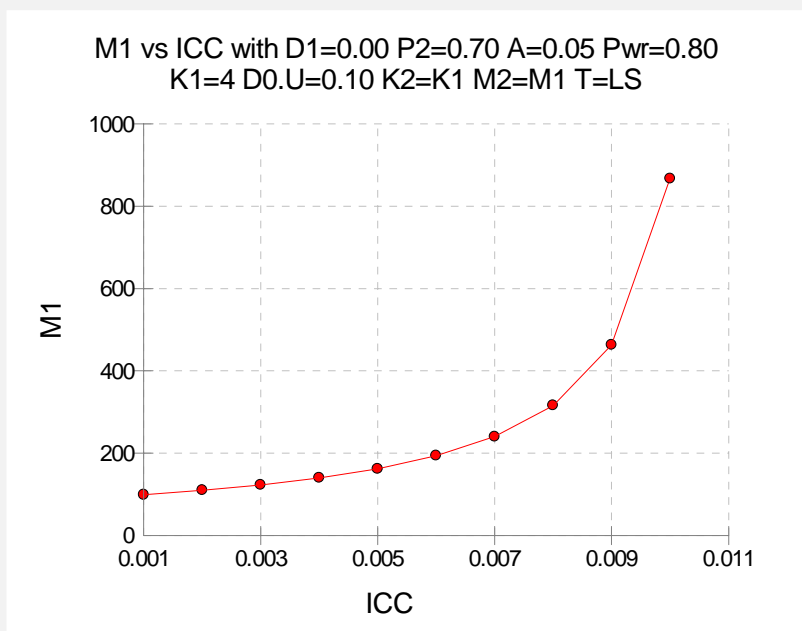
Numeric Results for Equivalence Tests Based on the Difference:  $P1 - P2$

$H_0: P1 - P2 \leq D0.L$  or  $P1 - P2 \geq D0.U$ .  $H_1: D0.L < P1 - P2 = D1 < D0.U$ .

Test Statistic: Score tests

	Group 1 Clusters/ Items	Group 2 Clusters/ Items	Intra- Cluster Corr. ICC	Prop Grp 2 P2	Lower Equiv. Grp 1 Prop P1.0L	Upper Equiv. Grp 1 Prop P1.0U	Lower Equiv. Margin Diff D0.L	Upper Equiv. Margin Diff D0.U	Actual Margin Diff D1	Alpha	Beta
Power	K1/M1	K2/M2									
0.8028	4/99	4/99	0.0010	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1972
0.8042	4/110	4/110	0.0020	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1958
0.8028	4/123	4/123	0.0030	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1972
0.8014	4/140	4/140	0.0040	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986
0.8014	4/162	4/162	0.0050	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986
0.8014	4/194	4/194	0.0060	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986
0.8014	4/240	4/240	0.0070	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986
0.8014	4/316	4/316	0.0080	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986
0.8014	4/463	4/463	0.0090	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986
0.8014	4/867	4/867	0.0100	0.7000	0.6000	0.8000	-0.1000	0.1000	0.0000	0.0500	0.1986

### Plots Section



The number of patients needed to be treated by each doctor ranges from 99 to 867 depending on the value of the intracluster correlation coefficient.

## Chapter 250

# Chi-Square Tests

## Introduction

The *Chi-square test* is often used to test whether sets of frequencies or proportions follow certain patterns. The two most common instances are tests of goodness of fit using multinomial tables and tests of independence in contingency tables.

The *Chi-square goodness of fit* test is used to test whether the distribution of a set of data follows a particular pattern. For example, the goodness-of-fit Chi-square may be used to test whether a set of values follow the normal distribution or whether the proportions of Democrats, Republicans, and other parties are equal to a certain set of values, say 0.4, 0.4, and 0.2.

The *Chi-square test for independence* in a contingency table is the most common Chi-square test. Here individuals (people, animals, or things) are classified by two (nominal or ordinal) classification variables into a two-way, contingency table. This table contains the counts of the number of individuals in each combination of the row categories and column categories. The Chi-square test determines if there is dependence (association) between the two classification variables. Hence, many surveys are analyzed with Chi-square tests.

The following table is an example of data arranged in a two-way contingency table. The rows of the table represent the stated political party of a respondent. The columns represent the respondent's answer to a question about whether they favor a certain proposition. The body of the table represents the number of individuals that fall into each cell (category). Note that the opinions of 311 individuals are recorded in this table.

(Count)	<u>Favor Proposition A</u>	
<u>Political Party</u>	<u>Yes</u>	<u>No</u>
Democrats	86	21
Republican	54	59
Others	34	57

The table below presents the row percentages for each category.

(Row Percentage)	<u>Favor Proposition A</u>	
<u>Political Party</u>	<u>Yes</u>	<u>No</u>
Democrats	80.4	19.6
Republican	47.8	52.2
Others	37.4	62.6

The Chi-square statistic tests whether the percentage of *Yes* responses remains constant across the three political parties. Notice that 80% of the Democrats said *Yes*, while only 37% of those in the Other category chose *Yes*. The Chi-square value for the above table is 5.59, which is statistically significant. Obviously, there is quite a shift in response pattern on this item across political parties.

## Effect Size

We begin by defining what we will call the *effect size*. For each cell of a table containing  $m$  cells, suppose there are two proportions considered: one specified by a null hypothesis and the other specified by an alternative hypothesis. Often, the proportions specified by the alternative hypothesis are those occurring in the data. Define  $p_{0i}$  to be the proportion in cell  $i$  under the null hypothesis and  $p_{1i}$  to be the proportion in cell  $i$  under the alternative hypothesis. The effect size,  $w$ , is calculated using the formula

$$w = \sqrt{\sum_{i=1}^m \frac{(p_{1i} - p_{0i})^2}{p_{0i}}}.$$

The formula for computing the Chi-square value,  $\chi^2$ , is

$$\begin{aligned}\chi^2 &= \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i} \\ &= N \sum_{i=1}^m \frac{(p_{0i} - p_{1i})^2}{p_{0i}},\end{aligned}$$

where  $N$  is the total count in all the cells. Hence, the relationship between  $w$  and  $\chi^2$  is

$$\chi^2 = Nw^2$$

Note that when you are dealing with a contingency table, the cell index,  $i$ , is often replaced by two indices, one representing columns and the other representing rows.

The effect size,  $w$ , was used by Cohen (1988) because it does not depend on the sample size. He sets a small value of  $w$  at 0.1, a medium value at 0.3, and a large value at 0.5. Although these are rather arbitrary settings, they are useful for planning purposes.

## Chi-Square Effect Size Estimator

*PASS* provides a special module to aid in finding an appropriate value for  $w$  called the Chi-Square Effect Size Estimator. This module may be loaded by pressing the *CS* button near the *W* (Effect Size) box or from the menus by selecting *PASS* and then *Other*.

You will find that as values are typed into the body of the table, the value of the effect size (shown at the bottom in the box labeled Effect Size -  $W$ ) is also changed. Using this utility program, you can quickly determine the impact of table configurations on the value of  $w$ .

For example, suppose the cell proportions under the null and alternative hypotheses are as follows:

Cell	1	2	3	4	
$p_{oi}$	0.25	0.25	0.25	0.25	(Null: Equal distribution across the four cells.)
$p_{li}$	0.40	0.20	0.20	0.20	(Alternative: cell 1 has twice the probability as the rest.)

To calculate  $w$ , first create the differences (ignoring the signs since the differences will be squared

Diff	0.15	0.05	0.05	0.05
------	------	------	------	------

Next, square the differences

Diff^2	0.0225	0.0025	0.0025	0.0025
--------	--------	--------	--------	--------

Divide by the null

X^2	0.09	0.01	0.01	0.01
-----	------	------	------	------

When these are summed, the result is 0.12. Taking the square of 0.12 gives the value of  $w$  as 0.3464.

As an experiment, load the Chi-Square Effect Size Estimator and enter these values on the Multinomial Test window. Enter the  $p_{oi}$  values in the column labeled Data Values and the  $p_{li}$  values in the column marked Hypothesized Proportions. Check that the value of  $w$  is 0.3464. Next, change the Data Values to 4,2,2,2 and the Hypothesized Proportions to 1,1,1,1. Check that the value of  $w$  is still the same.

## Calculating the Power

The power is calculated as follows:

1. Find  $x_\alpha$  such that  $1 - \chi^2(x_\alpha | df) = \alpha$ , where  $\chi^2(x_\alpha | df)$  is the area to the left of  $x$  under a Chi-square distribution with  $df$  degrees of freedom.
2. Power =  $1 - \chi'^2_{df, \lambda}$ , where  $\chi'^2_{k, \lambda}$  is the left-tail area of the noncentral Chi-square distribution with  $k$  degrees of freedom and noncentrality parameter  $\lambda$ . Note that  $\lambda = Nw^2$ .

## 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 *W*, *DF*, *N*, *Alpha*, and *Beta*. 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.

### DF (Degrees of Freedom)

This options specifies the degrees of freedom of the Chi-square test. For a test of independence in a contingency table, the degrees of freedom is  $(R-1)(C-1)$  where *R* is the number of rows and *C* is the number of columns. For example, for a 3-by-4 table,  $DF = (3-1)(4-1) = 6$ .

In a goodness of fit test, the degrees of freedom is the number of cells minus one. You may have to further adjust it for every distributional parameter that is estimated from the data. For example, suppose a Chi-square goodness-of-fit will be used to test the adequacy of the normality assumption on a set of 300 observations. Two parameters, the mean and variance, are estimated from the data. Suppose the data are categorized into six categories.  $DF = 6 - 2 - 1 = 3$ .

### W (Effect Size)

This is the value of *w*, the effect size. If you have Chi-square values that you want to analyze, use the following formula to transform them to *w*'s:

$$w = \sqrt{\frac{\chi^2}{N}}$$

Remember that a small value of *w* is 0.1, a medium value is 0.3, and a large value is 0.5.

### N (Sample Size)

This option specifies the number of individuals whose responses are recorded in the table. This number should be greater than or equal to the number of cells in the table.

## 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 row (or column) proportions when in fact they 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. 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 row proportions when in fact they are different.

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

## Example 1 - Finding the Power for an Existing Contingency Table

This example will compute the power of the Chi-square test of independence of the data in the contingency table that was discussed at the beginning of this chapter. If you would like to follow along, load the Chi-Square Effect Size Estimator window, select the Contingency Table tab, enter 86, 54, 34 in the first column and 21, 59, 57 in the second column. The results are Chi-square = 41.708829,  $DF = 2$ ,  $N = 311$ , and  $W = 0.366213$ .

We will compute the power when  $\alpha = 0.01, 0.05$ , and  $0.10$ . For evaluation purposes, we will compute the power when  $N = 20, 50, 100$ , and  $200$  as well as at  $311$ .

### 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
DF .....	<b>2</b>
N .....	<b>20 50 100 200 311</b>
W.....	<b>0.366213</b>
Alpha.....	<b>0.01 0.05 0.10</b>
Beta.....	<i>Ignored since this is the Find setting.</i>

### Annotated Output

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



## Numeric Results

### Numeric Results

Power	N	W	Chi-Square	DF	Alpha	Beta
0.12127	20	0.3662	2.6822	2	0.01000	0.87873
0.29104	20	0.3662	2.6822	2	0.05000	0.70896
0.41007	20	0.3662	2.6822	2	0.10000	0.58993
0.39621	50	0.3662	6.7056	2	0.01000	0.60379
0.63538	50	0.3662	6.7056	2	0.05000	0.36462
0.74622	50	0.3662	6.7056	2	0.10000	0.25378
0.78214	100	0.3662	13.4112	2	0.01000	0.21786
0.91678	100	0.3662	13.4112	2	0.05000	0.08322
0.95512	100	0.3662	13.4112	2	0.10000	0.04488
0.98840	200	0.3662	26.8224	2	0.01000	0.01160
0.99795	200	0.3662	26.8224	2	0.05000	0.00205
0.99927	200	0.3662	26.8224	2	0.10000	0.00073
0.99980	311	0.3662	41.7088	2	0.01000	0.00020
0.99998	311	0.3662	41.7088	2	0.05000	0.00002
1.00000	311	0.3662	41.7088	2	0.10000	0.00000

### 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.

W is the effect size--a measure of the magnitude of the Chi-Square that is to be detected.

DF is the degrees of freedom of the Chi-Square distribution.

Alpha is the probability of rejecting a true null hypothesis.

Beta is the probability of accepting a false null hypothesis.

### Summary Statements

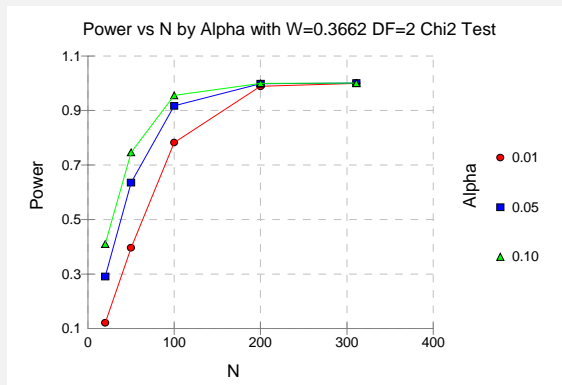
A sample size of 20 achieves 12% power to detect an effect size (W) of 0.3662 using a 2 degrees of freedom Chi-Square Test with a significance level (alpha) of 0.01000.

This report shows the values of each of the parameters, one scenario per row. The definitions of each column are given in the Report Definitions section.

Note that in this particular example, a reasonable power of about 0.80 is reached for all values of alpha once the sample size is greater than 100.

The values from this table are plotted in the chart below.

## Plots Section



This plot shows the relationship between sample size, power, and alpha.

## Example 2 - Finding the Sample Size

A survey is being planned that will contain several questions with three possible answers: agree, neutral, disagree. The researchers are planning to analyze the questionnaires using Chi-square tests of independence in two-way contingency tables. How many respondents are needed to detect small ( $w = 0.1$ ), medium ( $w = 0.3$ ), or large ( $w = 0.5$ ) effects if all hypothesis testing will be done at the 0.05 significance level?

Since the researchers are planning for 3-by-3 tables,  $DF = (3 - 1)(3 - 1) = 4$ .

### 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N</b>
DF .....	<b>4</b>
N .....	<i>Ignored since this is the Find setting.</i>
W .....	<b>0.1 0.3 0.5</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.10 0.20</b>

### Annotated Output

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

### Numeric Results

Numeric Results						
Power	N	W	Chi-Square	DF	Alpha	Beta
0.90010	1541	0.1000	15.4100	4	0.05000	0.09990
0.80018	1194	0.1000	11.9400	4	0.05000	0.19982
0.90157	172	0.3000	15.4800	4	0.05000	0.09843
0.80130	133	0.3000	11.9700	4	0.05000	0.19870
0.90198	62	0.5000	15.5000	4	0.05000	0.09802
0.80243	48	0.5000	12.0000	4	0.05000	0.19757

This report shows that for 80% power, 1194 (about 1200) respondents are needed to detect small effects, 133 respondents are needed to detect medium effects, and 48 respondents are needed to detect large effects.

## Example 3 - Validation using Cohen

Cohen (1988) page 251 presents an example in which  $W = 0.30$  and  $0.40$ ,  $N = 140$ ,  $\alpha = 0.01$ , and  $DF = 2$ . He gives the power as  $0.75$  for  $W = 0.3$  and  $0.97$  for  $W = 0.4$ .

### 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
DF .....	<b>2</b>
N.....	<b>140</b>
W.....	<b>0.3 0.4</b>
Alpha .....	<b>0.01</b>
Beta.....	<i>Ignored since this is the Find setting.</i>

### Annotated Output

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

### Numeric Results

Numeric Results						
Power	N	W	Chi-Square	DF	Alpha	Beta
0.74841	140	0.3000	12.6000	2	0.01000	0.25159
0.96641	140	0.4000	22.4000	2	0.01000	0.03359

*PASS* matches Cohen's power values of  $0.75$  and  $0.97$ .

## Example 4 - Finding the Sample Size for a Normality Goodness-of-Fit Test

A researcher is planning a study to determine if the distribution of scores on a certain test is normal. He plans to divide the test scores from his sample into five intervals of equal probability under the normal distribution using the sample mean and sample variance. After experimenting with the Chi-Square Effect Size Estimator, he decides that he must be able to detect a departure from normality of  $w = 0.20$ . He sets his significance level at 0.10 so that he will be lenient in his rejection of normality. He decides to focus on a power of 0.80. How large of a sample size will the researcher need?

The value of  $DF = 5 - 2 - 1 = 2$ , since there are five intervals and two parameters, mean and variance, are used.

### 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.

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N</b>
DF .....	<b>2</b>
N .....	<i>Ignored since this is the Find setting.</i>
W .....	<b>0.25</b>
Alpha .....	<b>0.10</b>
Beta .....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

Numeric Results						
Power	N	W	Chi-Square	DF	Alpha	Beta
0.80046	193	0.2000	7.7200	2	0.10000	0.19954

This report shows that for 80% power, 193 observations are needed.

## Chapter 260

# ROC Curve

Receiver operating characteristic (ROC) curves are used to assess the accuracy of a diagnostic test. The technique is used when you have a criterion variable which will be used to make a yes or no decision based on the value of this variable. The area under the ROC curve (AUC) is a popular summary index of an ROC curve.

This module computes power and sample size when a new diagnostic test is compared to an existing (gold) standard. Two approaches are available: the approach of Hanley and McNeil (1982) is used when the criterion variable is continuous and the approach of Obuchowski and McClish (1997) is used when the criterion variable is a discrete rating scale.

## Technical Details

In the following, we suppose that we have two groups of patients, those with a condition of interest (the positive group) and those without it (the negative group). This classification may be known from extensive diagnosis or based on the value of another diagnostic test. The diagnostic test of interest is performed on each patient and the resulting test value is recorded. At each specified cutoff value of the criterion variable, the true positive rate (TPR) and the false positive rate (FPR) are calculated. A plot of the TPR versus the FPR allows you study the consequences of using various cutoff values. This plot is called the *ROC curve*.

It should be noted that TPR is similar to the statistical power of the diagnostic test at a particular cutoff value of the criterion variable. Similarly, FPR is an estimate of the probability that the diagnostic test results in a type I ( $\alpha$ ) error. Thus the ROC curve may be interpreted as a plot of the diagnostic test's power versus its significance level at various possible criterion cutoff values.

Users of ROC curves have developed special names for TPR and FPR. They call TPR the *sensitivity* of the test and  $1 - \text{FPR}$  the *specificity* of the test. Statisticians will be more familiar with using the word *power* instead of sensitivity and the phrase ' $1 - \alpha$ ' instead of specificity.

An ROC curve may be summarized by the area under it (AUC). This area has an additional interpretation. Suppose that a rater is asked to study two subjects, one that is actually disease positive and one that is disease negative. The AUC is equal to the probability that the rater will give the disease positive subject a higher score than the disease negative subject. That is, the AUC is the probability that the rater will correctly order the two subjects as to which is more likely to have the disease.

Several methods of computing the AUC have been proposed. One method uses the trapezoidal rule to calculate the AUC directly. Another method, called the *binormal model*, computes the area by fitting two normal distributions to the data.

## The Binormal Model

Let  $X$  denote the distribution of the criterion variable for negative (normal) patients and  $Y$  denote the distribution of the criterion variable for positive (diseased) patients. It is assumed that

$$X \sim N(\mu_-, \sigma_-^2)$$

and

$$Y \sim N(\mu_+, \sigma_+^2)$$

For a particular cutoff value of the criterion variable,  $c$ , the true positive rate is given by

$$\begin{aligned} TPR(c) &= P(Y > c) \\ &= 1 - \Phi\left(\frac{c - \mu_+}{\sigma_+}\right) \\ &= \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right) \end{aligned}$$

where  $\Phi(z)$  is the cumulative normal distribution.

Similarly, the false positive rate is given by

$$\begin{aligned} FPR(c) &= P(X > c) \\ &= 1 - \Phi\left(\frac{c - \mu_-}{\sigma_-}\right) \\ &= \Phi\left(\frac{\mu_- - c}{\sigma_-}\right) \end{aligned}$$

The ROC curve is thus the curve traced out by the functions

$$[FPR(c), TPR(c)] = \left[ \Phi\left(\frac{\mu_- - c}{\sigma_-}\right), \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right) \right]$$

The area under the ROC curve, AUC, is defined as

$$\begin{aligned} \theta &= \int_{-\infty}^{\infty} TPR(c) FPR'(c) dc \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right) \phi\left(\frac{\mu_- - c}{\sigma_-}\right) \left(-\frac{1}{\sigma_-}\right) dc \\ &= \int_{-\infty}^{\infty} \Phi(A + Bv) \phi(v) dv \\ &= \Phi\left(\frac{A}{\sqrt{1 + B^2}}\right) \end{aligned}$$

where  $c = \mu_+ - v\sigma_-$ , and

$$A = \frac{|\mu_+ - \mu_-|}{\sigma_+}$$

$$B = \frac{\sigma_-}{\sigma_+}$$

Maximum likelihood estimates of  $A$  and  $B$  can be computed and used to compute AUC. The variances and covariance of these MLE's can be estimated from Fisher's information matrix.

Define  $\Delta = \theta_0 - \theta_1$  to be the difference in the accuracies (AUC's) of two diagnostic tests. A hypothesis test of whether the two AUC's are different amounts to testing whether  $\Delta = 0$ . The test statistic for this test is

$$Z = \frac{\hat{\theta} - \theta_1}{\sqrt{\text{var}_0(\hat{\theta})}}$$

where  $\text{var}_0(\hat{\theta})$  is the variance of  $\hat{\theta}$  under the null hypothesis of equality. The above test statistic gives the following formulae for computing sample size or power

$$N_+ = \frac{\left(z_\alpha \sqrt{V(\theta_0)} + z_\beta \sqrt{V(\theta_1)}\right)^2}{(\theta_1 - \theta_0)^2}$$

$$\beta = \Phi\left(\frac{|\theta_1 - \theta_0| \sqrt{N_+} - z_\alpha \sqrt{V(\theta_0)}}{\sqrt{V(\theta_1)}}\right)$$

## Rating Data

For a criterion variable yielding a discrete rating, Obuchowski (1998) recommends

$$V(\theta) = f^2 \left(1 + \frac{B^2}{R} + \frac{A^2}{2}\right) + g^2 B^2 \left(\frac{1+R}{2R}\right)$$

where

$$f = \frac{E_1}{\sqrt{2\pi(1+B^2)}}$$

$$g = -\frac{ABE_1}{\sqrt{2\pi(1+B^2)^3}}$$

$$E_1 = \exp\left(-\frac{A^2}{2+2B^2}\right)$$

The value of  $A$  can be found as

$$A = \Phi^{-1}(\theta) \sqrt{1+B^2}$$

For the most conservative results, Obuchowski (1998) recommends setting  $B = 1$ , so that

$$A = \Phi^{-1}(\theta)\sqrt{2}$$

## Continuous Data

For a criterion variable yielding a continuous result, Obuchowski (1998) suggests that the following formula of Hanley and McNeil (1983) is more appropriate

$$V(\theta) = \frac{\theta}{R(2 - \theta)} + \frac{2\theta^2}{1 + \theta} - \theta^2 \left( \frac{1 + R}{R} \right)$$

## 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

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

### AUC0 (Area under Curve)

Specify one or more values of the AUC for the diagnostic test. The range of values is from 0.5 (indicative of a test useless in diagnosis) to 1.0 (indicative of a test that is perfect in diagnosis).

Since the AUC may include a portion of the ROC curve that is not of interest because the FPR values are unrealistic, you may be interested in only a portion of the area. In this case, you can specify a range of FPR values for which the area is to be calculated. Unfortunately, the definition of the area becomes more difficult. When analyzing the whole ROC curve, the area is known to be between 0.50 and 1.0. Following the suggestion of Obuchowski and McClish (1997), the following transformation is applied so that the values of AUC remain between 0.5 and 1.0.

$$AUC' = \frac{1}{2} \left( 1 + \frac{AUC - \min}{\max - \min} \right)$$

where

$$\max = FPR2 - FPR1$$

$$\min = \frac{\max}{2} (FPR2 + FPR1)$$



Thus, when a partial range is entered for FPR1 and FPR2, the values entered here are assumed to be AUC' and are translated to AUC using the above formulas.

## AUC1

Specify one or more values of AUC under the alternative hypothesis. The range of values is from 0.5 (indicative of a test useless in diagnosis) to 1.0 (indicative of a test that is perfect in diagnosis). Note that, as discussed above, this is the value of AUC' when a partial area is being analyzed.

## 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 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. 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 when in fact it is false. Values must be between zero and one. Traditionally, the values from 0.05 to 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. 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.

## Type of Data

Specify the type of data that will be collected from the tests. The formulas for the variance are determined by this option. Possible types are:

**Continuous.** The test results are from a continuum of possible values. The Hanley and McNeil (1983) variance formulas are used. Note that this option does not allow a partial range of FPR values to be analyzed.

**Discrete.** The test results are from a small set of rating values such as 1, 2, 3, 4, 5. The Obuchowski & McClish (1997) variance formulas are used.

## Alternative Hypothesis

Specify whether the test is *one-sided* or *two-sided*. When a two-sided test is selected, the value of alpha is divided by two.

Note that most researchers assume that, unless stated otherwise, all statistical tests are two-sided. If you use a one-sided test, you should clearly state and justify this in all reports.

## Lower FPR

This option specifies the lower (left) limit of the false positive rate (FPR) for which the area is to be computed. If the area under the whole ROC curve is wanted, set this value to 0.0. If the partial area is wanted, set this value to the desired left limit.

Note that the range of possible values is  $0.0 \leq \text{Lower FPR} < \text{Upper FPR} \leq 1.0$

## Upper FPR

This option specifies the upper (right) limit of the false positive rate (FPR). If the area under the whole ROC curve is wanted, set this value to 1.0. If the partial area is wanted, set this value to the desired right limit.

Note that the range of possible values is  $0.0 \leq \text{Lower FPR} < \text{Upper FPR} \leq 1.0$

## B (SD Ratio)

$B$  is the ratio of the standard deviation of the negative group to the positive group ( $\sigma_-/\sigma_+$ ) for the diagnostic test. That is, assuming the binormal model

$$B = \frac{\sigma_-}{\sigma_+}$$

Note that this parameter is ignored for continuous data.

Although  $B$  can be any positive number, typical values are between 0.3 and 3.0. Obuchowski suggests that if the value of  $B$  is not known, a value of 1.0 is used since this will result in a conservative (extra large) sample size. She reports that in her experience, typical values are much less than 1.0, often near 0.3.

## N+ (Size of Positive Group)

Specify the number of patients, that is, the sample size, in the positive (abnormal or diseased) group. Note that these values are ignored when you are solving for  $N+$ . You may enter a range of values such as *10 to 100 by 10*.

## N- (Size of Negative Group)

Specify the number of patients, that is, the sample size, in the negative (normal) group. Enter *Use R* to base  $N-$  on the value of  $N+$ . You may enter a range of values such as *10 to 100 by 10*.

### Use R

When *Use R* is entered here,  $N-$  is calculated using the formula

$$N- = [R(N+)]$$

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  $N+ = N-$ , enter *Use R* here 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  $N-$  is set to *Use R*.

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

## Example1 - Calculating Power

An investigator wants to study the accuracy of a diagnostic test which yields measurements on a rating scale from 1 to 5. Historically, such tests have had an AUC of 0.80. The investigator wants to investigate three alternative AUC values: 0.825, 0.850, and 0.900. A two-sided test is planned with a significance level of 0.05. Since no other information is available, B is set to 1.0. The investigator would like to achieve a power of 90% in the study. Patients without the disease under study are about twice as frequent as patients with the disease. The investigator wants to see results for a sample size of up to 6000 patients.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
AUC1 .....	<b>0.80</b>
AUC2 .....	<b>0.825 0.85 0.9</b>
Alpha .....	<b>0.05</b>
Beta .....	<b><i>Ignored since this is the Find Setting</i></b>
Type of Data .....	<b>Discrete (Ratings)</b>
Alternative Hypothesis .....	<b>Two-Sided Test</b>
Lower FPR .....	<b>0.00</b>
Upper FPR .....	<b>1.00</b>
B .....	<b>1</b>
N+ .....	<b>20 50 100 250 500 1000 2000</b>
N- .....	<b>Use R</b>
R .....	<b>2</b>
<b>Axis Tab</b>	
Vertical Range .....	<b>User</b>
Minimum .....	<b>0</b>
Maximum .....	<b>1</b>
Number of Tick Marks .....	<b>10</b>

### Annotated Output

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

## Numeric Report

### Numeric Results for Testing $AUC_0 = AUC_1$ with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 1.0. B = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC0'	AUC1'	Diff'	AUC0	AUC1	Diff	Alpha	Beta
0.0481	20	40	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.9519
0.0739	50	100	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.9261
0.1146	100	200	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.8854
0.2365	250	500	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.7635
0.4321	500	1000	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.5679
0.7264	1000	2000	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.2736
0.9550	2000	4000	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.0450
0.0870	20	40	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.9130
0.1834	50	100	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.8166
0.3491	100	200	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.6509
0.7369	250	500	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.2631
0.9629	500	1000	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0371
0.9997	1000	2000	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0003
1.0000	2000	4000	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0000
0.2489	20	40	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.7511
0.6563	50	100	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.3437
0.9474	100	200	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0526
1.0000	250	500	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000
1.0000	500	1000	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000
1.0000	1000	2000	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000
1.0000	2000	4000	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000

### Report Definitions

Power is the probability of rejecting a false null hypothesis.

N+ is the sample size from the positive (diseased) population.

N- is the sample size from the negative (non-diseased) population.

Alloc Ratio is the Sample Allocation Ratio ( $R = N- / N+$ ).

AUC0' is the adjusted area under the ROC curve under the null hypothesis.

AUC1' is the adjusted area under the ROC curve under the alternative hypothesis.

Diff' is  $AUC1' - AUC0'$ . This is the adjusted difference to be detected.

AUC0 is the actual area under the ROC curve under the null hypothesis.

AUC1 is the actual area under the ROC curve under the alternative hypothesis.

Diff is  $AUC1 - AUC0$ . This is the difference to be detected.

Alpha is the probability of rejecting a true null hypothesis.

Beta is the probability of accepting a false null hypothesis.

FPR1, FPR2 are the lower and upper bounds on the false positive rates.

B is the ratio of the standard deviations of the negative and positive groups.

### Summary Statements

A sample of 20 from the positive group and 40 from the negative group achieve 5% power to detect a difference of 0.0250 between the area under the ROC curve (AUC) under the null hypothesis of 0.8000 and an AUC under the alternative hypothesis of 0.8250 using a two-sided z-test at a significance level of 0.0500. The data are discrete (rating scale) responses. The AUC is computed between false positive rates of 0.000 and 1.000. The ratio of the standard deviation of the responses in the negative group to the standard deviation of the responses in the positive group is 1.000.

This report shows the power for each of the sample sizes. Most of the definitions are standard. However, a special explanation must be given for AUC and AUC'.

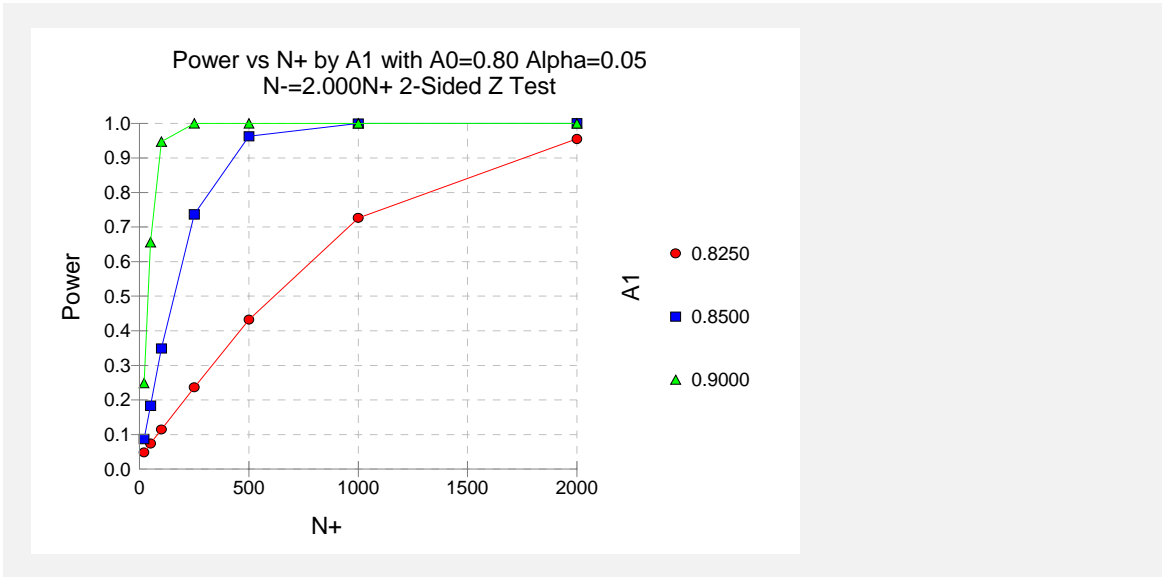
## AUC'

This is the adjusted area under the curve. A rescaling, discussed earlier, has been applied so that the minimum area is 0.5 and the maximum area is 1.0.

## AUC

This is the actual area under the curve. This value will equal the adjusted area when the FPR range is set from 0.0 to 1.0. Otherwise, these values will be different.

Plot Section



This plot shows the power versus the sample size for the three values of AUC1.

## Example2 - Calculating Sample Size

Continuing on with Example1, the investigator wants to know the exact sample size needed for each of the three values of AUC2. The investigator wants to look at the Numeric Report. The panel from Example1 is modified as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N+</b>
<b>Reports Tab</b>	
Show Definitions .....	<b>Unchecked</b>
Show Plots .....	<b>Unchecked</b>
Summary Statement Rows .....	<b>0</b>

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

### Numeric Report

#### Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 1.0. B = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC0'	AUC1'	Diff'	AUC0	AUC1	Diff	Alpha	Beta
0.9001	1582	3164	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.0999
0.9007	381	762	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0993
0.9024	85	170	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0976

This report shows the sample size needed to achieve 90% power for each value of AUC1.

## Example3 - Partial Area under Curve

Continuing on with Example2, the investigator knows that FPR values between 0.0 and 0.20 are the only values of interest. Hence, he wants to investigate the sample size needed when the FPR range is confined to this range.

The panel from Example2 is modified as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Upper FPR .....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

#### Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 0.200. B = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC0'	AUC1'	Diff'	AUC0	AUC1	Diff	Alpha	Beta
0.9001	2663	5326	0.8000	0.8250	0.0250	0.1280	0.1370	0.0090	0.0500	0.0999
0.9002	645	1290	0.8000	0.8500	0.0500	0.1280	0.1460	0.0180	0.0500	0.0998
0.9013	144	288	0.8000	0.9000	0.1000	0.1280	0.1640	0.0360	0.0500	0.0987

Note that the necessary sample size has almost doubled.

## Example4 - Validation using Obuchowski

The formulas used in this module were given in Obuchowski and McClish (1997). On page 1538, they provide an example which will be duplicated here. The study investigated the accuracy of MRI for detecting abnormalities in patients with symptomatic knees. In order to do this, they wanted to know the sample size that would be needed to construct a 95% confidence interval so that the length of the confidence interval is no more than 0.10.

The measure of diagnostic accuracy is the AUC from an FPR of 0.0 to an FPR of 1.0. The allocation ratio is 1.5.  $B = 1.0$ . The value of  $A$  is found to be 1.2. This translates to an  $AUC_0$  of 0.7995. The value of  $AUC_1 = AUC_0 + 0.10 / 2$ , where 0.10 is the maximum length of the confidence interval. A two-tailed confidence interval is envisioned in which  $\alpha$  is 0.05. In order to find the sample size of a confidence interval, the power is set to 50%. In their article, they found  $N_+ = 161$  and  $N_- = 242$ .

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N+</b>
AUC1 .....	<b>0.7995</b>
AUC2 .....	<b>0.8495</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.50</b>
Type of Data .....	<b>Discrete (Ratings)</b>
Alternative Hypothesis .....	<b>Two-Sided Test</b>
Lower FPR .....	<b>0.00</b>
Upper FPR .....	<b>1.00</b>
B .....	<b>1</b>
N+ .....	<b><i>Ignored since this is the Find Setting</i></b>
N- .....	<b>Use R</b>
R .....	<b>1.5</b>

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

### Numeric Results

#### Numeric Results for Testing $AUC_1 = AUC_2$ with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 0.200.  $B_1 = 1.000$ .  $B_2 = 1.000$ . Allocation Ratio = 2.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.5026	162	243	0.7995	0.8495	0.0500	0.7995	0.8495	0.0500	0.0500	0.4974

Note that the sample sizes of 162 and 243 are within one of the results of Obuchowski. The difference occurs because their values of 161 and 242 produce a power that is slightly less than 0.5, so PASS increased the sample size slightly.



## Chapter 265

# Comparing Two ROC Curves

Receiver operating characteristic (ROC) curves are used to summarize the accuracy of diagnostic tests. The technique is used when a criterion variable is available which is used to make a yes or no decision. The area under the ROC curve (AUC) is a popular summary index of an ROC curve.

This module computes power and sample size for comparing the AUC's of two diagnostic tests obtained from the same patients. The methodology of Obuchowski and McClish (1997) is used when the criterion variable yields a discrete value. The methodology of Hanley and McNeil (1983) is used when the criterion variable yields a continuous value.

## Technical Details

In the following, we suppose that we have two groups of patients, those with a condition of interest (the disease) and those without it. A patient's classification may be known from extensive diagnosis or based on the value of another diagnostic test. The diagnostic tests of interest are performed on each patient and the resulting test values are recorded. At each specified cutoff value of the criterion variable, the true positive rate (TPR) and the false positive rate (FPR) are calculated. An ROC curve is generated by plotting TPR versus FPR. The plot allows the consequences of using various cutoff values to be evaluated. The area under the ROC curve, either for the whole or partial range, is often used as a summary measure of the accuracy of the test.

It should be noted that TPR is similar to the statistical power of the diagnostic test at a particular cutoff value of the criterion variable. Similarly, FPR is an estimate of the probability that the diagnostic test results in a type I ( $\alpha$ ) error. Thus the ROC curve may be interpreted as a plot of the diagnostic test's power versus its significance level at various possible criterion cutoff values.

Users of ROC curves have developed special names for TPR and FPR. They call TPR the *sensitivity* of the test and  $1 - \text{FPR}$  the *specificity* of the test. Statisticians will be more familiar with using the word *power* instead of sensitivity and the phrase ' $1 - \alpha$ ' instead of specificity.

An ROC curve may be summarized by the area under it (AUC). This area has an additional interpretation. Suppose that a rater is asked to study two subjects, one that is actually disease positive and one that is disease negative. The AUC is equal to the probability that the rater will give the disease positive subject a higher score than the disease negative subject. That is, the AUC is the probability that the rater will correctly order the two subjects as to which is more likely to have the disease.

Several methods of computing the AUC have been proposed. One method uses the trapezoidal rule to calculate the AUC directly. Another method, called the *binormal model*, computes the area by fitting two normal distributions to the data.

## The Binormal Model

Let  $X$  denote the distribution of the criterion variable for normal (non-diseased) patients and  $Y$  denote the distribution of the criterion variable for abnormal (diseased) patients. It is assumed that

$$X \sim N(\mu_-, \sigma_-^2)$$

and

$$Y \sim N(\mu_+, \sigma_+^2)$$

The partial area under the ROC curve,  $AUC$ , is defined as

$$\theta_i = \int_{c_1}^{c_2} \Phi(A_i + B_i v) \phi(v) dv$$

where  $\Phi(z)$  is the cumulative normal distribution,  $c_j = \Phi^{-1}(FPR_j)$ , and

$$A_i = \frac{\mu_{i+} - \mu_{i-}}{\sigma_{i+}}$$

$$B_i = \frac{\sigma_{i-}}{\sigma_{i+}}$$

Note that for the full range area under the curve,  $c_1 = -\infty$  and  $c_2 = \infty$ .

Maximum likelihood estimates of  $A$  and  $B$  can be computed. The variances and covariance of these MLE's can be estimated from Fisher's information matrix.

Define  $\Delta = \theta_1 - \theta_2$  to be the difference in the accuracies (AUC's) of the two tests. A test of whether the two AUC's are different amounts to testing whether  $\Delta = 0$ . The test statistic for this test is

$$Z = \frac{\hat{\Delta} - 0}{\sqrt{\text{var}_0(\hat{\Delta})}}$$

where  $\text{var}_0(\hat{\Delta})$  is the variance of  $\hat{\Delta}$  under the null hypothesis of equality. The above test statistic results in the following formula for computing sample size

$$N_+ = \frac{\left( z_\alpha \sqrt{V_0(\hat{\Delta})} + z_\beta \sqrt{V_{Alt}(\hat{\Delta})} \right)^2}{\Delta^2}$$

## Rating Data

When the criterion values are discrete rating values, Obuchowski and McClish (1997) showed that the variances could be calculated using

$$V_0(\hat{\Delta}) = V(\hat{\theta}_1) + V(\hat{\theta}_1) - 2C(\hat{\theta}_1, \hat{\theta}_1)$$

$$V_{Alt}(\hat{\Delta}) = V(\hat{\theta}_1) + V(\hat{\theta}_2) - 2C(\hat{\theta}_1, \hat{\theta}_2)$$

where

$$V(\hat{\theta}_i) = f_i^2 \left( 1 + \frac{B_i^2}{R} + \frac{A_i^2}{2} \right) + g_i^2 \left( B_i^2 \left( \frac{1+R}{2R} \right) \right)$$

$$C(\hat{\theta}_1, \hat{\theta}_2) = f_1 f_2 \left( r_+ + r_- \frac{B_1 B_2}{R} + r_+^2 \frac{A_1 A_2}{2} \right) + \frac{g_1 g_2 B_1 B_2 (r_-^2 + R r_+^2)}{2R} + \frac{f_1 g_2 A_1 B_2 r_+^2}{2} + \frac{f_2 g_1 A_2 B_1 r_+^2}{2}$$

$$f_i = \frac{E_{1i} E_{3i}}{\sqrt{2\pi E_{2i}}}$$

$$g_i = \frac{E_{1i} E_{4i}}{\sqrt{2\pi E_{2i}}} - \frac{A_i B_i E_{1i} E_{3i}}{\sqrt{2\pi E_{2i}^3}}$$

$$E_{1i} = \exp\left(-\frac{A_i^2}{2 + 2B_i^2}\right)$$

$$E_{2i} = 1 + B_i^2$$

$$E_{3i} = \Phi(c_2) - \Phi(c_1)$$

$$E_{4i} = \exp\left(-\frac{c_1^2}{2}\right) - \exp\left(-\frac{c_2^2}{2}\right)$$

$$c_j = \frac{\Phi^{-1}(FPR_j) + \frac{A_j B_j}{1 + B_j^2}}{\sqrt{1 + B_j^2}}$$

$$R = \frac{N_-}{N_+}$$

$$A_i = B_i \Phi^{-1}(TNR_i) - \Phi^{-1}(FPR_i)$$

$r_-$  and  $r_+$  are the correlations between the results of the two diagnostics tests for normal and abnormal patients, respectively. For the most conservative results, set  $B_i = 1$ .

## Continuous Data

When the criterion values are continuous, Obuchowski (1998) suggests that the following formulas of Hanley and McNeil (1983) are more appropriate. Note that these formulas cannot be used for evaluating the AUC for a partial range.

$$V(\hat{\Delta}) = V(\hat{\theta}_1) + V(\hat{\theta}_2) - 2C(\hat{\theta}_1, \hat{\theta}_2)$$

where

$$V(\hat{\theta}_i) = \frac{\theta_i}{R(2 - \theta_i)} + \frac{2\theta_i^2}{1 + \theta_i} - \theta_i^2 \left( \frac{1 + R}{R} \right)$$

$$C(\hat{\theta}_1, \hat{\theta}_2) = 2r\sqrt{V(\theta_1)V(\theta_2)}$$

and  $r$  is derived from a special table provided by Hanley and McNeil (1983).

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

## AUC1 (Area under Curve)

Specify one or more values of the AUC for diagnostic test 1. The range of values is from 0.5 (indicative of a test useless in diagnosis) to 1.0 (indicative of a test that is perfect in diagnosis).

Since the AUC may include a portion of the ROC curve that is not of interest because the FPR values are unrealistic, you may be interested in only a portion of the area. In this case, you can specify a range of FPR values for which the area is to be calculated. Unfortunately, the definition of the area becomes more difficult. When analyzing the whole ROC curve, the area is known to be between 0.50 and 1.0. Following the suggestion of Obuchowski and McClish (1997), the following transformation is applied so that the values of AUC remain between 0.5 and 1.0.

$$AUC' = \frac{1}{2} \left( 1 + \frac{AUC - \min}{\max - \min} \right)$$

where

$$\max = FPR2 - FPR1$$

$$\min = \frac{\max}{2}(FPR2 + FPR1)$$

Thus, when a partial range is entered for FPR1 and FPR2, the values entered here are assumed to be AUC' and are translated to AUC using the above formulas.

## AUC2 (Area under Curve)

Specify one or more values of the AUC for diagnostic test 2. The range of values is from 0.5 (indicative of a test useless in diagnosis) to 1.0 (indicative of a test that is perfect in diagnosis). Note that, as discussed above, this is the value of AUC' when a partial area is being analyzed.

## 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 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. 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 when in fact it is false. Values must be between zero and one. Traditionally, the values from 0.05 to 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. 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.

## Type of Data

Specify the type of data that will be collected from the tests. The formulas for the variance are determined by this option. Possible types are:

**Continuous.** The test results are from a continuum of possible values. The Hanley and McNeil (1983) variance formulas are used. Note that this option does not allow a partial range of FPR values to be analyzed.

**Discrete.** The test results are from a small set of rating values such as 1, 2, 3, 4, 5. The Obuchowski & McClish (1997) variance formulas are used.

## Alternative Hypothesis

Specify whether the test is *one-sided* or *two-sided*. When a two-sided test is selected, the value of alpha is divided by two.

Note that most researchers assume that, unless stated otherwise, all statistical tests are two-sided. If you use a one-sided test, you should clearly state and justify this in all reports.

## Lower FPR

This option specifies the lower (left) limit of the false positive rate (FPR) for which the area is to be computed. If the area under the whole ROC curve is wanted, set this value to 0.0. If the partial area is wanted, set this value to the desired left limit.

Note that the range of possible values is  $0.0 \leq \text{Lower FPR} < \text{Upper FPR} \leq 1.0$

## Upper FPR

This option specifies the upper (right) limit of the false positive rate (FPR). If the area under the whole ROC curve is wanted, set this value to 1.0. If the partial area is wanted, set this value to the desired right limit.

Note that the range of possible values is  $0.0 \leq \text{Lower FPR} < \text{Upper FPR} \leq 1.0$

## B1 (SD Ratio)

$B1$  is the ratio of the standard deviation of the negative group to the positive group ( $SD_-/SD_+$ ) for diagnostic test 1. That is, assuming the binormal model

$$B1 = \frac{\sigma_{1-}}{\sigma_{1+}}$$

Note that this parameter is ignored for continuous data.

Although  $B1$  can be any positive number, typical values are between 0.3 and 3.0. Obuchowski suggests that if the value of  $B1$  is not known, a value of 1.0 is used since this will result in a conservative (extra large) sample size. She reports that in her experience, typical values are much less than 1.0, often near 0.3.

## B2 (SD Ratio)

$B2$  is the ratio of the standard deviation of the negative group to the positive group ( $SD_-/SP_+$ ) for diagnostic test 2. That is, assuming the binormal model

$$B2 = \frac{\sigma_{2-}}{\sigma_{2+}}$$

Note that this parameter is ignored for continuous data.

Although  $B2$  can be any positive number, typical values are between 0.3 and 3.0. Obuchowski suggests that if the value of  $B2$  is not known, a value of 1.0 is used since this will result in a conservative (extra large) sample size. She reports that in her experience, typical values are much less than 1.0, often near 0.3.

## Correlation+

This is the correlation between the two diagnostic-test scores for the positive group. Although correlations can range between -1 and 1, typical values are from 0.3 to 0.6.

Note that if you want to analyze a design in which a separate set of patients receive each diagnostic test, this may be done by setting this correlation value to 0.

## Correlation-

This is the correlation between the two diagnostic-test scores for the negative group. Although correlations can range between -1 and 1, typical values are from 0.3 to 0.6.

Note that if you want to analyze a design in which a separate set of patients receive each diagnostic test, this may be done by setting this correlation value to 0.

## **N+ (Size of Positive Group)**

Specify the number of patients, that is, the sample size, in the positive (abnormal or diseased) group. Note that these values are ignored when you are solving for  $N+$ . You may enter a range of values such as *10 to 100 by 10*.

## **N- (Size of Negative Group)**

Specify the number of patients, that is, the sample size, in the negative (normal) group. Enter *Use R* to base  $N-$  on the value of  $N+$ . You may enter a range of values such as *10 to 100 by 10*.

### **Use R**

When *Use R* is entered here,  $N-$  is calculated using the formula

$$N- = [R(N+)]$$

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  $N+ = N-$ , enter *Use R* here 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  $N-$  is set to *Use R*.

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

## Example1 - Calculating Power

An investigator wants to compare the accuracy of two diagnostic tests which yield measurements on a rating scale from 1 to 5. Historically, such tests have had an AUC of 0.80. The investigator wants to investigate three alternative AUC values: 0.825, 0.850, and 0.900. A two-sided test is planned with a significance level of 0.05. Historically, both the positive and negative correlations between the responses on two such tests have been close to 0.60. Since no other information is available, B1 and B2 are both set to 1.0. The investigator would like to achieve a power of 90% in the study. Patients without the disease under study are about twice as frequent as patients with the disease. The investigator wants to see results for a sample size of up to 6000 patients.

Using these values, the investigator fills out the PASS: ROC Curve - Two Tests panel as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>Beta and Power</b>
AUC1 .....	<b>0.80</b>
AUC2 .....	<b>0.825 0.85 0.9</b>
Alpha .....	<b>0.05</b>
Beta .....	<b><i>Ignored since this is the Find Setting</i></b>
Type of Data .....	<b>Discrete (Ratings)</b>
Alternative Hypothesis .....	<b>Two-Sided Test</b>
Lower FPR .....	<b>0.00</b>
Upper FPR .....	<b>1.00</b>
B1 .....	<b>1</b>
B2 .....	<b>1</b>
Correlation+ .....	<b>0.6</b>
Correlation- .....	<b>0.6</b>
N+ .....	<b>20 50 100 250 500 1000 2000</b>
N- .....	<b>Use R</b>
R .....	<b>2</b>
<b>Axis Tab</b>	
Vertical Range .....	<b>User</b>
Minimum .....	<b>0</b>
Maximum .....	<b>1</b>
Number of Tick Marks .....	<b>10</b>

### Annotated Output

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



## Numeric Report

### Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 1.0. B1 = 1.000. B2 = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.0501	20	40	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.9499
0.0733	50	100	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.9267
0.1084	100	200	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.8916
0.2104	250	500	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.7896
0.3744	500	1000	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.6256
0.6426	1000	2000	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.3574
0.9090	2000	4000	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.0910
0.0920	20	40	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.9080
0.1737	50	100	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.8263
0.3083	100	200	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.6917
0.6442	250	500	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.3558
0.9116	500	1000	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0884
0.9969	1000	2000	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0031
1.0000	2000	4000	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0000
0.2470	20	40	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.7530
0.5494	50	100	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.4506
0.8496	100	200	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.1504
0.9978	250	500	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0022
1.0000	500	1000	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000
1.0000	1000	2000	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000
1.0000	2000	4000	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0000

### Report Definitions

Power is the probability of rejecting a false null hypothesis.

N+ is the sample size from the positive (diseased) population.

N- is the sample size from the negative (non-diseased) population.

Alloc Ratio is the Sample Allocation Ratio ( $R = N- / N+$ ).

AUC1' is the adjusted area under the ROC curve for diagnostic test 1.

AUC2' is the adjusted area under the ROC curve for diagnostic test 2.

Diff' is  $AUC2' - AUC1'$ . This is the adjusted difference to be detected.

AUC1 is the actual area under the ROC curve for diagnostic test 1.

AUC2 is the actual area under the ROC curve for diagnostic test 2.

Diff is  $AUC2 - AUC1$ . This is the difference to be detected.

Alpha is the probability of rejecting a true null hypothesis.

Beta is the probability of accepting a false null hypothesis.

FPR1, FPR2 are the lower and upper bounds on the false positive rates.

B1 and B2 are the ratios of the standard deviations of the negative and positive groups for each test.

### Summary Statements

A sample of 20 from the positive group and 40 from the negative group achieve 5% power to detect a difference of 0.0250 between a diagnostic test with an area under the ROC curve (AUC) of 0.8000 and another diagnostic test with an AUC of 0.8250 using a two-sided z-test at a significance level of 0.0500. The data are discrete (rating scale) responses. The AUC is computed between false positive rates of 0.000 and 1.000. The ratio of the standard deviation of the responses in the negative group to the standard deviation of the responses in the positive group for diagnostic test 1 is 1.000 and for diagnostic test 2 is 1.000. The correlation between the two diagnostic tests is assumed to be 0.600 for the positive group and 0.600 for the negative group.

This report shows the power for each of the sample sizes. Most of the definitions are standard. However, a special explanation must be given for AUC and AUC'.

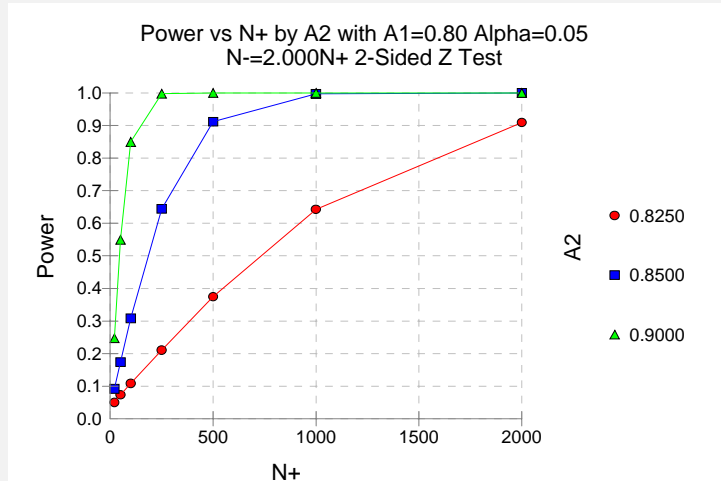
## AUC'

This is the adjusted area under the curve. A rescaling, discussed earlier, has been applied so that the minimum area is 0.5 and the maximum area is 1.0.

## AUC

This is the actual area under the curve. This value will equal the adjusted area when the FPR range is set from 0.0 to 1.0. Otherwise, these values will be different.

## Plot Section



This plot shows the power versus the sample size for the three values of AUC1.

## Example2 - Calculating Sample Size

Continuing Example1, the investigator wants to know the exact sample size needed for each of the three values of AUC2. The investigator wants to look at the Numeric Report. The panel from Example1 is modified as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	N+
<b>Reports Tab</b>	
Show Definitions .....	Unchecked
Show Plots .....	Unchecked
Summary Statement Rows .....	0

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

### Numeric Report

#### Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 1.0. B1 = 1.000. B2 = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.9001	1937	3874	0.8000	0.8250	0.0250	0.8000	0.8250	0.0250	0.0500	0.0999
0.9002	480	960	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.0500	0.0998
0.9012	117	234	0.8000	0.9000	0.1000	0.8000	0.9000	0.1000	0.0500	0.0988

This report shows the sample size needed to achieve 90% power for each value of AUC2.

## Example3 - Partial Area Under Curve

Continuing Example2, the investigator knows that FPR values between 0.0 and 0.20 are the only values of interest. Hence, he wants to investigate the sample size needed when the FPR range is confined to this range.

The panel from Example2 is modified as follows.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Upper FPR .....	<b>0.20</b>

### Annotated Output

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

### Numeric Results

#### Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 0.200. B1 = 1.000. B2 = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.9000	4095	8190	0.8000	0.8250	0.0250	0.1280	0.1370	0.0090	0.0500	0.1000
0.9002	1012	2024	0.8000	0.8500	0.0500	0.1280	0.1460	0.0180	0.0500	0.0998
0.9001	242	484	0.8000	0.9000	0.1000	0.1280	0.1640	0.0360	0.0500	0.0999

Note that the necessary sample size has more than doubled.

## Example4 - Validation using Obuchowski

The formulas used in this module were given in Obuchowski and McClish (1997). On pages 1538 - 1540, they provide an example which will be duplicated here. The study compared an automated classification system with an expert mammographer in their ability to find malignant breast lesions. The measure of diagnostic accuracy is the AUC from an FPR of 0.0 to an FPR of 0.2. The allocation ratio is 2.  $B1 = B2 = 1.0$ .  $\text{Correlation}+ = \text{Correlation}- = 0.6$ . The values of A1 and A2 are found to be 2.6 and 1.9. These translate to adjusted AUC's of 0.922222 and 0.819444. A two-tailed test is envisioned in which alpha is 0.05. A power of 80% is desired. In their article, they found  $N+ = 109$  and  $N- = 218$ .

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N+</b>
AUC1 .....	<b>0.80</b>
AUC2 .....	<b>0.825 0.85 0.9</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.20</b>
Type of Data .....	<b>Discrete (Ratings)</b>
Alternative Hypothesis .....	<b>Two-Sided Test</b>
Lower FPR .....	<b>0.00</b>
Upper FPR .....	<b>0.20</b>
B1 .....	<b>1</b>
B2 .....	<b>1</b>
Correlation+ .....	<b>0.6</b>
Correlation- .....	<b>0.6</b>
N+ .....	<b><i>Ignored since this is the Find Setting</i></b>
N- .....	<b>Use R</b>
R .....	<b>2</b>

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

### Numeric Results

#### Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0. FPR2 = 0.200. B1 = 1.000. B2 = 1.000. Allocation Ratio = 2.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.8027	109	218	0.9222	0.8194	-0.1028	0.1720	0.1350	-0.0370	0.0500	0.1973

Note that the sample sizes of 109 and 218 match exactly with the results of Obuchowski.

## Example5 - Validation using Hanley

The formulas for continuous data were given in Hanley and McNeil (1982). On page 34 of their article they provide a table of sample sizes calculated using their formulas. We will duplicate their results for  $AUC1 = 0.70$  and  $AUC2 = 0.75$ . Using a one-sided test of significance with  $\alpha = 0.05$  and a sample allocation ratio of 1.0, they found the number of subjects for both the positive and negative groups to be 652, 897, and 1131 for statistical powers of 80%, 90%, and 95%, respectively.

When using Hanley and McNeil's formulation, the values of B1, B2, FPR1, and FPR2 are ignored. Also, in this case, the correlations are set to 0.0.

### Setup

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

<u>Option</u>	<u>Value</u>
<b>Data Tab</b>	
Find .....	<b>N+</b>
AUC1 .....	<b>0.7</b>
AUC2 .....	<b>0.75</b>
Alpha .....	<b>0.05</b>
Beta .....	<b>0.2 0.1 0.05</b>
Type of Data .....	<b>Continuous</b>
Alternative Hypothesis .....	<b>One-Sided Test</b>
Correlation+ .....	<b>0.0</b>
Correlation- .....	<b>0.0</b>
N+ .....	<b><i>Ignored since this is the Find Setting</i></b>
N- .....	<b>Use R</b>
R .....	<b>1</b>

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

## Numeric Results

### Numeric Results for Testing AUC1 = AUC2 with Continuous Data

Test Type = One-Sided. FPR1 = 0.0. FPR2 = 1.0. B1 = 1.000. B2 = 1.000. Allocation Ratio = 1.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.9501	1129	1129	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0499
0.9001	897	897	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0999
0.8003	652	652	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.1997

Note that the sample sizes of 897 and 652 match exactly with the results of Hanley and McNeil. The 1129 is two less than their 1131. This difference may be due to refinements in computing the normal probability distribution used in PASS. You can compare these sample sizes by calculating their power.

## Numeric Results

### Numeric Results for Testing AUC1 = AUC2 with Continuous Data

Test Type = One-Sided. FPR1 = 0.0. FPR2 = 1.0. B1 = 1.000. B2 = 1.000. Allocation Ratio = 1.000.

Power	N+	N-	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha	Beta
0.9499	1128	1128	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0501
0.9501	1129	1129	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0499
0.9502	1130	1130	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0498
0.9504	1131	1131	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0496
0.9505	1132	1132	0.7000	0.7500	0.0500	0.7000	0.7500	0.0500	0.0500	0.0495

**Chapter 365** Note that the power for 1129 is 0.9501 while the power for 1131 is 0.9505. This is only a slight difference and explains why this value showed up in their table.

# 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-Sundugchi, 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 Processes*. 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.

## References-2

---

- 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. *Graphical 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 a 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.
- Dunnnett, 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.

## References-6

---

- 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." *Annals 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 superseded 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 follow and provides lots of examples.
- Jolliffe, I.T.** 1972. "Discarding variables in a principal component analysis, I: Artificial 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. Pharmacokinetic. 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, 2<sup>nd</sup> 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 I*. Van Nostrand Reinhold. 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 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 consistent 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 contained 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. At 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, Massachusetts.
- 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](http://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.

- 2x2 cross-over: mean difference, 500-1; non-inferiority, 510-1; ratio, 505-1
- 3D parameters, 28
- 3D tab, 28
- Abbreviations tab, 31
- Abort, 18
- Accrual time: log-rank, 705-5
- All-contrasts power, 590-4
- All-pair power, 580-5, 585-4
- Alpha, 44; adjusting, 60; MANOVA, 605-10; multiple comparisons, 580-2, 585-2, 590-1
- Alpha spending function: means, 475-1; proportions, 220-1; survival, 710-1
- Alternative hypothesis, 42, 55
- Analysis of variance: fixed effects ANOVA, 560-1; one-way, 550-1; randomized block, 560-1
- ANOVA: factorial, 560-1; fixed effects, 560-1; multiple comparisons, 575-2, 580-3; multiple contrasts, 590-2; one-way, 550-1; randomized block, 560-1; repeated measures, 570-1; simulation, 555-1; three-way, 560-1; two-way, 560-1
- Any-contrast power, 590-4
- Any-pair power, 580-5, 585-4
- : ANOVA, 550-2; Mann-Whitney, 430-4; MANOVA, 605-1; repeated measures, 570-2; t-test, 430-4; , 410-3, 410-4, 410-5
- AUC: ROC curve, 260-2; ROC curves, 265-1, 265-5
- Autocorrelation, 570-27
- Average absolute deviation: standard deviation estimator, 905-3
- Axes tab, 27
- Axis: horizontal, 23
- Axis color, 28
- Axis maximum, 27
- Axis minimum, 27
- Background tab, 30
- Balaam's design, 530-1, 530-2, 535-1, 535-2, 540-2
- Bar Chart Options, 23
- : , 105-8
- Best: multiple comparisons, 575-5
- Beta, 45; log-rank survival test, 700-5; , 105-12; , 410-7; , 410-6, 410-10; show power, 23; simulating a, 630-3; , 410-6, 410-10
- Between standard deviation, 570-15
- Bimodal data: simulating, 630-23
- Binomial, 43, 100-2; proportion, 115-1; simulating a, 630-5
- Binomial tests, 100-4
- Binormal model: ROC curve, 260-2; ROC curves, 265-2
- Bioequivalence: proportions, 215-1
- Blackwelder: risk ratio, 205-29
- Blocks, 560-9
- Bonferroni test, 590-3
- Bootstrap test: simulation, 410-4, 490-5, 495-6
- Brownian motion: group sequential, 475-5; proportions, 220-5; survival, 710-6
- Cage edge, 29
- Cage flip, 29
- Cage wall, 29
- Carryover effects, 570-3
- Case-control: matched, 255-1
- Cauchy: simulating a, 630-5
- Cell means, 560-1
- Chart Type, 23
- Charts: interactive formatting, 25
- Chi-square: estimator, 900-1; one variance, 650-1
- Chi-square test, 250-1; proportions, 200-6
- Circularity, 570-14
- Clinical trial: three-stage, 130-1
- Clinical trials, 120-2
- Cluster randomization: equivalence, 240-1; means, 480-1; proportions, 230-1; proportions, 240-1; non-inferiority, 235-1
- Clusters: cluster randomization, 480-3
- Cochran-Mantel-Haenszel, 225-1
- Coefficient alpha, 815-1, 1
- Coefficient of variation, 905-10; cross-over, 505-3; mean ratio, 445-2, 455-3, 470-4, 515-4, 535-4
- Cohort study, 135-1
- Color: axis, 28; grid lines, 27
- Color max, 29
- Color min, 29
- Color of legend, 26
- Color palette, 29
- Comparisons: multiple comparisons, 575-1; one-way, 550-1
- Comparisonwise: error rate, 580-2, 585-2, 590-1
- Compound symmetry, 570-14
- Confidence coefficient: mean, 420-2, 420-3; proportion, 115-2
- Confidence interval: mean, 420-1; proportion, 115-1
- Constant: simulating a, 630-6
- Consumer's risk, 405-1
- Contingency table, 250-1; chi-square, 900-1
- Continuity correction, 100-5; proportions, 220-8; two proportions, 205-8
- Contrast coefficients, 550-7
- Contrast matrix, 570-11
- Contrasts, 550-1, 590-1; MANOVA, 605-6; multiple, 590-2
- Control: multiple comparisons, 575-2, 585-1
- Controlled variables: multiple regression, 865-5

## 2 Index

---

- Correlated proportions: equivalence, 165-1; McNemar test, 150-1
- Correlation: cluster randomization, 480-3; intraclass, 810-1; linear regression, 855-1; one, 800-1; two, 805-1
- Correlation coefficient, 800-1
- Correlation tolerance: paired means using simulation, 490-11
- Covariance: Hotelling's T<sup>2</sup>, 600-5; MANOVA, 605-11
- Covariance matrix, 570-14
- Covariate: logistic regression, 860-6
- Cox regression, 850-1
- Creating data: simulation, 630-1
- Cronbach's alpha, 815-1, 1
- Coefficient of variation, 525-4, 545-5
- Equivalence tests, 540-3
- Hypothesis, 545-3
- Cross-over: Balaam's design, 540-2; equivalence, 520-1, 525-1, 540-1, 545-1; higher order, 535-1, 540-1; higher-order, 530-1; mean difference, 500-1; non-inferiority, 510-1, 515-1, 530-1, 535-1; ratio, 505-1, 525-1, 535-1; repeated measures, 570-46
- Cubic: contrast, 550-7
- Customizing toolbars, 13
- Data, 38; simulation of, 630-1
- Data Tab, 21
- Decimals, 22
- Default: template, 16
- Depth, 28
- Diagnostic testing: ROC curve, 260-1
- Difference: equivalence, 110-4, 165-3, 460-1, 520-1, 540-1; non-inferiority, 105-4, 160-3, 450-1, 530-1; proportion, 100-6; proportions, 52, 200-3, 205-3, 210-4, 215-4
- Discordant pairs, 150-2
- Distribution: combining, 630-16; mixing, 630-16
- Distributions: simulating paired, 490-2; simulation, 630-1
- Documentation, 19, 39
- Donner & Klar: cluster randomization, 230-2
- Double exponential, 400-7
- Drift: group sequential, 475-5; proportions, 220-5
- Dunn's test, 590-3; power, 590-4
- Dunnett's test, 585-3; multiple comparisons, 575-3
- Dunnett's test, 585-1
- Dunn's test: power, 590-1
- Edit menu, 36
- Effect size, 46; ANOVA, 560-3, 560-7; chi-square, 900-1, 900-2; chi-square test, 250-2; multiple regression, 865-3; one-way ANOVA, 550-3; randomized block, 560-2
- Effect Size: one-way ANOVA, 550-13
- Balaam's design, 545-2
- Equivalence: cluster randomization, 240-1; correlated proportions, 165-1; cross-over, 520-1, 525-1, 540-1, 545-1; difference, 460-1; hypothesis, 215-4, 215-5; limits, 495-8; margin, 450-9; Mean, 520-1; mean ratio, 470-1; means, 465-1; paired means, 495-1; proportion, 110-1; proportions, 165-1, 215-1; ratio, 165-6; ratio, 470-1; simulation, 495-3; t-test, 465-3
- Equivalence hypothesis, 50
- Equivalence margin: non-inferiority, 455-4
- Error rates: multiple comparison, 580-2, 585-2, 590-1
- Errors, 42
- Exit, 18
- Experimentwise: error rate, 580-2, 585-2, 590-1
- Exponential: log-rank, 705-3; means, 435-1; simulating a, 630-7
- Exponential data: simulation, 410-28
- Exponential mean, 405-1
- Exponential test: simulation, 410-5
- Exposed, 255-5
- Exposure, 870-1
- Exposure probability: matched case-control, 255-3
- F: simulating a, 630-8
- Factor: fixed, 560-5; random, 560-5
- Factorial ANOVA, 560-1
- Familywise: error rate, 580-2, 585-2, 590-1
- Farrington - Manning test: difference, 210-11; equivalence, 215-11; non-inferiority, 210-11; ratio, 215-11
- Farrington-Manning test: two proportions, 205-11
- File menu, 34
- File Menu, 17
- File Name: template file, 32
- Files: template, 32
- : , 105-12
- Finite population correction: t test, 400-6
- Finite population size, 420-2; proportion, 115-2
- Fisher's Exact test: proportions, 200-5
- Fisher-z transformation, 805-1
- Fixed factor, 560-5
- Fleming: one-stage design, 120-1
- Folders, 1
- Follow-up: log-rank, 705-4
- Fonts: changing, 37
- Format menu, 37, 40
- Formatting: charts interactively, 25
- FPR: ROC curve, 260-6; ROC curves, 265-1, 265-6
- F-test: Geisser-Greenhouse, 570-4; one-way, 555-2; simulation, 555-1; two variances, 655-1
- Games-Howell: multiple comparison, 580-1
- Games-Howell test, 580-4
- Gamma: simulating, 630-24; simulating a, 630-9
- Gart - Nam test: difference, 210-13; equivalence, 215-12; non-inferiority, 210-13; ratio, 215-12
- Gart-Nam test: two proportions, 205-12
- Geisser-Greenhouse, 570-1
- Geisser-Greenhouse F-test, 570-4
- General linear multivariate model: MANOVA, 605-2
- General Linear Multivariate Model, 570-3
- Generating data, 630-1
- Goodness of fit, 250-1; chi-square, 900-1
- Grid color, 27
- Grid line style, 27
- Grid lines, 27
- Group Sample Size, 555-5
- Group sample size pattern, 580-10
- Group sequential test: log-rank, 710-1; means, 475-1; proportions, 220-1; survival, 710-1
- Hazard rate: Cox regression, 850-1

- Hazard rates: log-rank, 700-2, 705-2
- Hazard ratio: group sequential, 710-2
- Help menu, 39
- Help system, 5
- Cross-over, 545-1
- Home window, 11
- Horizontal Axis, 23
- Hotelling's T2, 600-1
- Hotelling-Lawley trace, 570-7, 605-1; MANOVA, 605-5
- Hotelling-Lawley trace, 570-1
- Hypergeometric, 100-2
- Hypotheses: ANOVA, 560-4; non-inferiority, 415-2; offset proportions, 205-5; superiority, 415-2; types, 48
- Hypothesis: difference, 210-4, 215-4; equivalence, 50, 110-4, 165-3, 215-4, 215-5; inequality, 48; introduction, 42; means, 55; non-inferiority, 49, 105-4, 105-5, 160-3, 160-6, 450-2; odds ratio, 110-6, 210-5, 215-5; , 410-7; one variance, 650-3; ratio, 110-5, 210-5, 215-5; superiority, 50, 105-4; Superiority, 210-4
- Hypothesis testing: introduction, 42
- Hypothesized mean, 550-14
- Hypothesized means, 550-6; randomized block, 560-4
- Icons, 12, 13
- Incidence rate, 135-2
- Independence test, 250-1
- Inequality: 2x2 cross-over, 500-1; correlated proportions, 150-1; hypothesis, 48; mean ratio, 445-1; proportion, 100-1; proportions, 200-1, 205-1
- Installation, 1
- Interactive Charts, 25
- Interactive Format, 25
- Intercept: linear regression, 855-1
- Interim analysis: means, 475-1; proportions, 220-1; survival, 710-1; three-stage, 130-1
- Intraclass correlation, 810-1
- Intraclass correlation: cluster randomization, 230-2, 480-3; cluster randomization, 235-3
- Isometric, 29
- Iterations: maximum, 22
- Kruskal-Wallis: multiple comparisons, 580-1; simulation, 555-1
- Kruskal-Wallis test, 580-4, 585-4; multiple comparisons, 585-1; simulation, 555-3
- Labels of plots, 26
- Lachin: log-rank test, 705-1
- Lan-DeMets: means, 475-1; proportions, 220-1; survival, 710-1
- Latin square: ANOVA, 560-19
- Legend, 23, 26
- Legend color, 26
- Likelihood ratio test: proportions, 200-8
- Likert-scale: simulating, 630-22; simulating a, 630-10; simulation, 410-23
- Line Chart options, 24
- Linear: contrast, 550-7
- Linear model, 570-3
- Linear model: ANOVA, 560-2
- Linear regression, 855-1; correlation, 800-1
- Load template, 32
- Log: cross-over, 515-3, 525-3, 545-4; mean ratio, 445-2, 455-2
- Log file, 34
- Log transformation: cross-over, 505-2, 535-4; ratio, 470-3
- Logistic, 400:7
- Logistic regression, 860-1
- Logit: logistic regression, 860-2
- Log-rank: group sequential test, 710-1
- Log-rank test, 700-1, 705-1
- Log-rank Non-Inferiority test, 705-15
- Mann-Whitney test, 430-1, 430-18; equivalence, 465-6; equivalence, 465-1; non-inferiority, 450-6; simulation, 440-4
- MANOVA, 605-1
- Mantel Haenszel test: proportions, 200-7
- Mantel-Haenszel, 225-1
- Margin of equivalence, 455-4; difference, 450-9
- Matched case-control, 135-2, 255-1
- Max time: sequential survival, 710-9
- Maximum: on axis, 27
- Maximum Iterations, 22
- McNemar test, 150-1
- Mean: confidence interval, 420-1; cross-over, 505-1; equivalence, 470-1, 525-1; exponential, 405-1; non-inferiority, 515-1; simulation, 410-1
- Mean difference: 2x2 cross-over, 500-1; cross-over, 510-1
- Mean ratio: equivalence, 545-1; inequality, 445-1; non-inferiority, 455-1
- Means: contrasts, 590-1; cross-over, 540-1; equivalence, 460-1, 465-1; exponential, 435-1; group sequential test, 475-1; hypothesized, 550-6; introduction, 55; MANOVA, 605-1; multiple comparisons, 585-1; non-inferiority, 450-1, 530-1; one-way, 555-1; paired, 490-1, 495-1; simulation, 59; simulation, 440-1
- Means matrix: MANOVA, 605-5
- Means matrix, 570-7
- Measurement error, 560-2
- Menu: edit, 36; file, 34; format, 37; help, 39; view, 37; window, 38
- Menus, 12, 17; file, 17
- Miettinen - Nurminen test: difference, 210-9, 215-9; equivalence, 215-9, 215-10; non-inferiority, 210-9; proportions, 205-9; Ratio, 215-10
- Minimum: on axis, 27
- Minimum detectable difference: multiple comparisons, 575-10; one-way ANOVA, 550-19; t-test, 400:13; two-sample t test, 430-15
- Monte Carlo, 57, 630-1
- MTBF: exponential mean, 405-1
- Multinomial: chi-square, 900-4; simulating a, 630-10
- Multiple comparisons, 575-1; Dunnett's test, 585-1; Games-Howell, 580-1; pair-wise, 580-1; power, 580-5, 585-4
- Multiple contrasts, 590-1; power, 590-4
- Multiple regression, 865-1
- Navigator, 38
- NCSS: quitting, 18

## 4 Index

---

- New Template, 17
- Nominal alpha: group sequential test of means, 475-13
- Noncentrality: one-way ANOVA, 550-3
- noncentrality parameter: one-way ANOVA, 550-4
- Non-inferiority: correlated proportions, 160-1; cross-over, 510-1, 515-1, 530-1; difference, 160-3; hypotheses, 450-2; hypotheses, 415-2; log-rank test, 705-15; mean difference, 450-1; mean ratio, 455-1; odds ratio, 210-5; paired means, 415-1, 490-8, 490-18; paired means, 415-2; proportion, 105-1; ; cluster randomization, 235-1; proportions, 210-1; ratio, 160-6, 210-5; z test, 210-7
- Non-Inferiority: simulation, 440-6
- Non-inferiority hypothesis, 49
- Non-Inferiority test: simulation, 440-17
- Non-inferiority tests, 510-3, 530-3, 535-3
- Non-null: proportions, 205-1
- Nonparametric: Mann-Whitney, 430-8; t-test, 400:7; Wilcoxon test, 400:1
- Normal: contaminated, 630-21; simulating, 630-19; simulating a, 630-11
- Nuisance parameter, 56; correlated proportions, 160-7, 165-7
- Nuisance parameters, 47
- Null case: proportions, 200-1
- Null hypothesis, 42, 55
- O'Brien-Fleming: means, 475-1; proportions, 220-1; survival, 710-1
- Odds ratio: equivalence, 110-6; logistic regression, 860-2; Mantel-Haenszel, 225-2; matched case-control, 255-1; McNemar test, 150-3; non-inferiority, 210-5; non-inferiority, 105-6; proportion, 100-7; proportions, 53, 200-3, 205-4, 215-5
- Odds ratio estimator, 910-1
- Offset: proportions, 205-1
- One-way ANOVA, 550-1
- Open Template, 17
- Options tab, 22
- Outliers: multiple comparisons, 580-26; simulation, 440-18, 555-19
- Outline window, 14
- Output: word processor, 33
- P value, 44
- Paired designs, 415-1
- Paired means: equivalence, 495-1; simulation, 490-1
- Paired proportions: non-inferiority, 160-1
- Paired t-test, 490-1
- Paired t-test, 400:14; assumptions, 400:3; non-inferiority, 415-1
- Paired t-tests, 400:1
- Pairwise comparisons: multiple comparisons, 575-8
- Pair-wise comparisons, 580-1
- Panel, 15
- PASS: starting, 7
- PASS Home, 11
- Password, 19, 39
- Patient entry: log-rank, 705-3
- PDF files, 19, 39
- Perspective, 28, 29
- Phase I trials, 120-2
- Phase II trials, 120-2
- Phi: matched case-control, 255-5
- Pillai-Bartlett trace, 570-6, 605-1; MANOVA, 605-4
- Pillai-Bartlett trace, 570-1
- Planned Comparisons, 550-1
- Plot Setup tab, 23
- Plot Text tab, 26
- Pocock: means, 475-1; proportions, 220-1; survival, 710-1
- Poisson: incidence, 135-1; simulating a, 630-11
- Poisson regression, 870-1
- Population size: t-test, 400:6
- Post-marketing surveillance, 135-1
- Power, 45; introduction, 41; means, 55; multiple comparisons, 580-5, 585-4; multiple contrasts, 590-4
- Prevalence: correlated proportions, 160-2, 165-2
- Print, 35
- Procedure Window, 15
- Producer's risk, 405-1
- Projection method, 28
- Proportion: confidence interval, 115-1; difference, 100-6; equivalence, 110-1; inequality, 100-1; non-inferiority, 105-1; odds ratio, 100-7; ratio, 100-6
- Proportional hazards regression, 850-1
- Proportions: Chi-square test, 200-6; cluster randomization, 240-1; inequality, 230-1; cluster randomization, 235-1; comparing, 51; correlated, 160-1, 165-1; difference, 52; equivalence, 215-1; Farrington - Manning test, 210-11; Fisher's exact, 200-5; Gart - Nam test, 210-13; group sequential test, 220-1; independent, 200-1; inequality, 200-1, 205-1; interim analysis, 220-1; interpretation, 54; introduction, 51; logistic regression, 860-1; matched case control, 255-1; McNemar test, 150-1; Miettinen - Nurminen test, 210-9, 215-9; non-inferiority, 160-1, 210-1; odds ratio, 53; odds ratio estimator, 910-1; offset, 205-1; paired, 160-1; ratio, 52; single-stage, 120-1; stratified, 225-1; superiority, 210-1; three-stage, 130-1; two-stage, 125-1
- Quadratic: contrast, 550-7
- Quitting, 35
- Random factor, 560-5
- Random number pool size, 495-13; paired means using simulation, 490-11
- Random numbers, 580-6, 585-5, 590-5, 630-1
- Randomized block ANOVA, 560-1
- Range on axis, 27
- Rating data: ROC curve, 260-3
- Ratio: cross-over, 505-1; equivalence, 110-5, 165-6; Farrington - Manning test, 215-11; Gart - Nam test, 215-12; inequality, 445-1; means, 445-1; Miettinen - Nurminen test, 215-10; non-inferiority, 160-6, 210-5, 455-2, 515-1, 535-1; proportion, 100-6, 105-5; proportions, 52, 200-3, 205-4, 215-5
- Ratios: equivalence, 470-1
- Regression: Cox, 850-1; linear, 855-1; logistic, 860-1; multiple, 865-1; Poisson, 870-1
- Rejection region, 44
- Repeated measures, 570-1

- Reports tab, 22
- Risk ratio: Blackwelder, 205-29; equivalence, 165-6
- ROC curve, 260-1
- ROC curves, 265-1
- Rotation of tickmarks, 26
- R-squared, 865-1; added, 865-5; logistic regression, 860-8
- RTF, 35
- RTF files, 33
- Ruler, 37
- Run menu, 18
- Sample size: introduction, 41
- Save template, 18, 32
- Score test: equivalence, 215-9, 215-10, 215-11, 215-12; Farrington - Manning test, 210-11, 215-11; Gart - Nam test, 210-13, 215-12; Miettinen - Nurminen test, 210-9, 215-9, 215-10; non-inferiority, 210-9, 210-11, 210-13; proportions, 205-9
- Sensitivity: correlated proportions, 160-2, 165-2; ROC curve, 260-1
- Serial numbers, 19, 39
- Show Beta as Power, 23
- Show tick marks, 28
- Sign test: simulation, 410-4, 490-5, 495-5
- Significance level, 44; adjusting, 60; multiple comparisons, 580-2, 585-2, 590-1
- Simon: two-stage, 125-1
- Simulation, 57, 630-1; equivalence, 465-1, 495-1; means, 440-1; multiple comparisons, 580-1, 580-5, 585-1, 585-5; multiple contrasts, 590-1, 590-5; one mean, 410-1; one-way, 555-1; paired means, 490-1, 495-1; random number generation, 580-6, 585-5, 590-5; size, 58; syntax, 630-16
- Single-stage design, 120-1
- Skewed data: one-way, 555-22; simulation, 410-14, 410-17, 440-20
- Skewed distribution: simulating a, 630-13
- Slope: linear regression, 855-1
- Specificity: correlated proportions, 160-2, 165-2; ROC curve, 260-1
- Spending functions: means, 475-2; proportions, 220-2
- Standard deviation, 56; estimator, 905-1; interpretation, 905-1; means, 550-13; one, 650-1; two, 655-1
- Standard deviation, 400:4
- Starting PASS, 3, 7
- Stratified designs, 225-1
- Student's T: simulating a, 630-12
- Style: grid line, 27
- Summary Statements, 22
- Superiority: hypotheses, 450-2; proportion, 105-4; proportions, 210-1, 210-4
- Superiority hypothesis, 50
- Superiority tests, 510-3, 530-3, 535-3
- Support, 5, 6
- Surface Chart options, 24
- Survival: log-rank, 700-1, 705-1
- Symbols, 30
- Symbols tab, 30
- System requirements, 1
- T: simulating a, 630-12
- Tab: abbreviations, 31
- Tabs, 21; axes, 27; data, 21; options, 22; plot setup, 23; reports, 22; symbols, 30; template, 32
- Tech support, 5, 6
- Template, 15; load, 32; save, 32
- Template Files, 32
- Template Id, 32
- Template tab, 32
- Templates, 17; automatic, 16; default, 16; loading, 17; new, 17; save, 18; saving, 17
- Test statistics, 47
- Text output, 22
- Thin Walls, 29
- Three-stage design, 130-1
- Tick marks, 27; show, 28
- Tickmark rotation, 26
- Titles of plots, 26
- Toolbar, 13, 20, 37, 40
- Treatment versus control: multiple comparisons, 575-2
- Trimmed t-test: equivalence, 465-5, 465-6; simulation, 440-3
- T-test: assumptions, 400:3; cluster randomization, 480-1; cross-over, 500-4, 510-4, 520-4, 540-4; equivalence, 215-8, 460-2, 465-3; equivalence, 520-4; equivalence, 540-4; non-inferiority, 210-8, 415-1, 415-5, 450-2, 450-5, 510-4; proportions, 200-8, 215-8; simulation, 410-1, 490-4, 495-4; simulation, 440-1; simulation, 440-2
- T-tests: assumptions, 430-4; one mean, 400:1; paired, 400:1; two means, 430-1
- Tukey-Kramer: simulation, 580-1
- Tukey-Kramer test, 580-3; multiple comparisons, 575-8
- Tukey's lambda: simulating a, 630-13
- : , 235-8, 235-11, 235-13, 235-15
- Two-sample t-test, 430-1; simulation, 440-2
- Two-Stage design, 125-1
- Type-I error, 42
- Type-II error, 42
- Uniform: simulating a, 630-14
- : , 235-12; repeated measures, 570-53; ROC Curves, 265-12; t-test, 400:17; two-sample t-test, 430-20
- Variance: one, 650-1
- Variances: two, 655-1
- Vertical viewing angle, 28
- View menu, 37
- Viewing angle: horizontal, 28; vertical, 28
- Wall color, 29
- Weibull: simulating a, 630-15
- Welch test: power, 590-4
- Welch's test: equivalence, 465-1; simulation, 440-3
- Welch's t-test: non-inferiority, 450-5
- Wilcoxon test, 400:7, 400:15, 415-8, 450-11; assumptions, 400:3; non-inferiority, 415-5; paired, 415-1; simulation, 410-3, 410-21, 490-1, 490-4, 495-5
- Wilcoxon test, 400:1
- Wilks' Lambda, 570-6, 605-1; MANOVA, 605-3
- Wilks' Lambda, 570-1
- Window menu, 38

## 6 Index

---

Winsorized test: equivalence, 465-5

Within standard deviation, 570-15

Within-subjects design: repeated measures, 570-2

Word processor: built in, 33

Z test: equivalence, 215-7; non-inferiority, 210-7;  
proportions, 205-7, 215-7

Z test - proportion -equivalence, 110-8

Z tests, 100-4

Zeros: two proportions, 205-19