User's Guide V

Tabulation, Item Analysis, Proportions, Diagnostic Tests, and Survival / Reliability

NCSS
Statistical System

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

NCSS User's Guide V

Dr. Jerry L. Hintze
Kaysville, Utah 84037

Direct inquiries to:

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

NCSS License Agreement

Important: The enclosed Number Cruncher Statistical System (NCSS) is licensed by Dr. Jerry L. Hintze to customers for their use only on the terms set forth below. Purchasing the system indicates your acceptance of these terms.

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

2. **COPYRIGHT.** NCSS and its documentation are copyrighted. You may not copy or otherwise reproduce any part of NCSS or its documentation, except that you may load NCSS 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.** NCSS may be backed up by you for your use on the same machine for which NCSS was purchased.

4. **RESTRICTIONS ON USE AND TRANSFER.** The original and any backup copies of NCSS and its documentation are to be used only in connection with a single computer. You may physically transfer NCSS from one computer to another, provided that NCSS is used in connection with only one computer at a time. You may not transfer NCSS electronically from one computer to another over a network. You may not distribute copies of NCSS or its documentation to others. You may transfer this license together with the original and all backup copies of NCSS and its documentation, provided that the transferee agrees to be bound by the terms of this License Agreement. Neither NCSS nor its documentation may be modified or translated without written permission from Dr. Jerry L. Hintze.

*You may not use, copy, modify, or transfer **NCSS**, 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.** Dr. Jerry L. Hintze does not and cannot warrant the performance or results that may be obtained by using NCSS. Accordingly, NCSS 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 NCSS is assumed by you. Should NCSS prove defective, you (and not Dr. Jerry L. Hintze nor his dealers) assume the entire cost of all necessary servicing, repair, or correction.

6. **LIMITED WARRANTY ON CD.** To the original licensee only, Dr. Jerry L. Hintze warrants the medium on which NCSS 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 NCSS is delivered. If, during this ninety-day period, a defect in a CD should occur, the CD may be returned to Dr. Jerry L. Hintze at his address, or to the dealer from which NCSS was purchased, and NCSS will replace the CD without charge to you, provided that you have sent a copy of your receipt for NCSS. 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, Dr. Jerry L. Hintze 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 Dr. Jerry L. Hintze nor anyone else who has been involved in the creation, production, or delivery of NCSS 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 NCSS 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 NCSS 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 NCSS and documentation together with all copies, modifications, and merged portions in any form.

9. **YOUR USE OF NCSS 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

Number Cruncher Statistical System (**NCSS**) is an advanced, easy-to-use statistical analysis software package. The system was designed and written by Dr. Jerry L. Hintze over the last several 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 (95, 98, ME, 2000, NT, 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.

Statistics is a broad, rapidly developing field. Updates and additions are constantly being made to the program. If you would like to be kept informed about updates, additions, and corrections, please send your name, address, and phone number to:

User Registration
NCSS
329 North 1000 East
Kaysville, Utah 84037

or Email you name, address, and phone number to:

Sales@NCSS.COM

NCSS maintains a website at **WWW.NCSS.COM** where we make the latest edition of NCSS 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.

NCSS maintains the following program and documentation copying policy. Copies are limited to a one person / one machine basis for "BACKUP" purposes only. You may make as many backup copies as you wish. Further distribution constitutes a violation of this copy agreement and will be prosecuted to the fullest extent of the law.

NCSS is not "copy protected." You may freely load the program onto your hard disk. We have avoided copy protection in order to make the system more convenient for you. Please honor our good faith (and low price) by avoiding the temptation to distribute copies to friends and associates.

We believe this to be an accurate, exciting, easy-to-use system. If you find any portion that you feel needs to be changed, please let us know. Also, we openly welcome suggestions for additions to the system.

User's Guide V
Table of Contents

User's Guide I
Table of Contents

User's Guide II
Table of Contents

User's Guide III
Table of Contents

User's Guide IV
Table of Contents

Chapter 500

Frequency Tables

Introduction

This procedure produces tables of frequency counts and percentages for discrete and continuous variables. This procedure serves as a summary reporting tool and is often used to analyze survey data. This procedure also calculates multinomial chi-square tests.

Frequency Tables

Frequency tables are generally produced on individual variables. For discrete data, the table records the number of observations (the frequency) for each unique value of the variable. For continuous data, you must specify a set of intervals (or bins). The frequency table now records the number of observations falling in each interval.

Frequency tables are useful for analyzing discrete data and for screening data for data entry errors.

Types of Categorical Variables

Note that we will refer to two types of categorical variables: *By* and *Break*. *Break* variables are used to split a database into subgroups. A separate set of reports is generated for each unique set of values of the *Break* variables. The values of a *By* variable are used to define the rows of the frequency table.

Data Structure

The data below are a subset of the real estate sales (RESALE) database provided with the software. This data gives the selling price, the number of bedrooms, the total square footage (finished and unfinished), and the size of the lots for 150 residential properties sold during the last four months in two states. Only the first 8 of the 150 observations are displayed.

RESALE dataset (subset)

State	Price	Bedrooms	TotalSqft	LotSize
Nev	260000	2	2042	10173
Nev	66900	3	1392	13069
Vir	127900	2	1792	7065
Nev	181900	3	2645	8484
Nev	262100	2	2613	8355

Missing Values

Missing values may be ignored or included in the table's counts and percentages.

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tabs

This panel specifies the variables that will be used in the analysis.

Specify one or more categorical variables whose categories (values) will appear along the rows of the frequency table. If more than one categorical variable is specified, a separate table will be generated for each variable.

Four types of categorical variables may be specified:

1. Variables containing text values.
2. Variables containing numeric values that are to be treated individually. For example, you might have used a set of index numbers like "1 2 3 4" to represent four states.
3. Variables containing numeric values that are to be grouped or combined into a set of predefined intervals. You specify the interval boundaries. For example, a variable containing age values might be grouped as "Under 21, 21 to 55, and Over 55." The key is that you specify the intervals.
4. Variables containing numeric values that are to be combined into a set of computer-generated intervals. You specify only the number of intervals. The program determines a set of equal-length intervals based on the minimum and maximum found in the data.

Variables for Use in Frequency Tables

Discrete Variables

This option specifies those variables that contain text and numeric values that are to be treated as discrete variables (Types 1 or 2). Variables containing text values are always listed here. Variables containing numeric values are listed here if you want each unique value to be treated separately.

Numeric Variables (Width)

Use this option to specify variables that contain numeric values that are to be combined into a set of computer-generated intervals (Type 4). The intervals are specified in the three boxes: Number, Minimum, and Width. Note that you can specify one, two, or all three of these options.

Number

The number of intervals to be created. If not enough intervals are specified to reach the maximum data value, more intervals are added.

Minimum

The minimum value or the left boundary of the first interval. This value must be less than the minimum data value.

Width

This is the width of an interval. A data value X is in this interval if Lower Limit < X <= Upper Limit. If this is left blank, it is calculated from the Number, Minimum, and maximum data value.

Numeric Variables (W)

This specifies those variables that contain numeric values that are to be combined into a set of user-specified intervals (Type 3). The interval boundaries are specified as a list in the Interval Upper Limits box.

Interval Upper Limits

Specify the upper limits of the intervals, separated by blanks or commas. For example, you would enter “1 3 5” to specify the four intervals: Under 1, 1 to 3, 3 to 5, and Over 5.

The logic structure of the interval is:

Lower Bound < *Value* <= *Upper Bound.*

Note that a “1” would be included in the “Under 1” interval, not the “1 to 3” interval. Also, a “5” would be included in the “3 to 5” interval, not the “Over 5” interval.

Frequency Variable

Frequency Variable

This optional variable specifies the number of observations that each row represents. When omitted, each row represents a single observation. If your data is the result of previous summarization, you may want certain rows to represent several observations. Note that negative values are treated as a zero weight and are omitted. Fractional values may be used.

This option is used when you want to enter previously tabulated data for a multinomial test.

Breaks Tab

This panel lets you specify up to eight break variables.

Select Break (Grouping) Variables

Break Variables

Specify one or more categorical variables whose distinct values will cause separate reports to be generated. Note that a separate set of reports (tables and plots) is generated for each unique set of values of these variables. Do not confuse these variables with the *Discrete Variables*, which specify the variables whose values will appear along the rows of a particular table.

Missing Tab

This panel lets you specify up to five missing values (besides the default of blank). For example, ‘0’, ‘9’, or ‘NA’ may be missing values in your database.

Missing Value Options

Missing Values

Specify up to five missing values here.

Missing Value Inclusion

Specifies whether to include observations with missing values in the tables.

- **Delete All**

  Indicates that you want the missing values totally ignored.

- **Include in Counts**

  Indicates that you want the number of missing values displayed, but you do not want them to influence any of the percentages.

- **Include in All**

  Indicates that you want the missing values treated just like any other category. They will be included in all percentages and counts.

Format Tab

The following options control the format of the reports.

Format Options

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want the table to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Label Justification

This option specifies whether the labels should be right or left justified above each column.

Data Justification

This option specifies whether the data should be right or left justified in each cell.

Split Column Headings

This option lets you select whether to split the column headings into two headings instead of one.

Tabs

These options let you specify the tab settings across the table. The output ruler is also modified by the settings of *Label Justification* and *Data Justification*.

First

Specifies the position of the first cell in inches. Note that the left-hand label always begins at 0.5 inches. Hence, the distance between this tab and 0.5 is the width provided for the row label information.

Maximum

Specifies the right border of the table. The number of tabs is determined based on *First*, the *Increment*, and this option. If you set this value too large, your table may not be printed correctly.

Increment

Specifies the width of a cell in inches.

Offset

The amount (inches) of offset to the right used with a decimal tab on a custom ruler so the data is aligned properly under the left-justified column labels.

Decimal Places

These options let you specify the number of decimal places used in the various items of the table.

By Labels

Specifies the number of decimal places displayed in the numeric categorical variable values. Note that *All* displays a single-precision (seven place accuracy).

Counts ... Table %s

Specifies the number of decimal places displayed in each statistic. Note that *All* displays the default amount.

Reports Tab

These options control which of the available reports and plots are displayed.

Select Reports

Frequency Table Report ... Table Percentages Report

Check each report that you want displayed.

Select Plots

Counts Plot ... Table Percentages Plot

Check each plot that you want displayed.

Combined Report

Show Combined Report

Specify whether to display this report.

Combined Report – Items on Report

Counts ... Table Percents

Specify whether to display these items on the combined report.

Combined Report – Options

Double Space Report

This option adds a blank row after each set of percentages in combined tables.

Multinomial Tab

These options let you specify a multinomial test on each frequency table.

Multinomial Test Options

Multinomial Test

Indicate whether to calculate a multinomial test

Expected Values for Multinomial Test

The multinomial test is a test of the proportions associated with a frequency table. In order to run the test, you need a set of hypothesized proportions. This box lets you enter these proportions. There are two ways of specifying the hypothesized proportions.

First, you can leave it blank to indicate that you want to test whether the proportions in the frequency table are all equal. This is the same as entering “1,1,1,1,1,1” (if the variable has six or fewer categories).

Second, you can enter a set of values separated by commas. These values are then rescaled so that they sum to one. For example, the entry “1,1,1,1” would be rescaled to “0.25, 0.25, 0.25, 0.25.” If you enter more numbers than are needed for a particular table, the extras are ignored. If you do not enter enough numbers, the remaining proportions are set to zero.

Plot Options Tab

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Connect Line(s)

Specifies whether connect the points with lines for easier interpretation of trends.

Plot Settings – Legend

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the appropriate default value.

Titles

Plot Title

This is the text of the title. The characters *{Y}*, *{X}*, and *{G}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Show Break as Title

Specifies whether the current values of any *Break* variables should be displayed as a second title line in the plot.

Symbols Tab

Specify the symbols used for each of the groups on the plots.

Plotting Symbols

Group 1-15

Specify the symbol used to designate a particular group. Double-click on a symbol or click on the button to the right of a symbol to specify the symbol's size, type, and color.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Standard Frequency Tables

The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Frequency Tables window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Frequency Tables window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Frequency Tables**. The Frequency Tables procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Frequency Tables window, select the **Variables tab**. (This is the default.)
- Double-click in the **Discrete Variables** text box. This will bring up the variable selection window.
- Select **State** and **City** from the list of variables and then click **Ok**. "State-City" will appear in the Discrete Variables box.

4 Specify the report format.

- Click on the **Format tab**.
- In **Variable Names**, select **Labels**.
- In **Value Labels**, select **Value Labels**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Frequency Table Output

Frequency Distribution of State

State	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
Nevada	88	88	58.67	58.67	||||||||||||||||||||
Virginia	62	150	41.33	100.00	|||||||||||||||

Frequency Distribution of City

Community	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
Silverville	27	27	18.00	18.00	|||||||
Los Wages	49	76	32.67	50.67	|||||||||||||
Red Gulch	12	88	8.00	58.67	||||
Politicville	27	115	18.00	76.67	|||||||
Senate City	24	139	16.00	92.67	||||||
Congresstown	11	150	7.33	100.00	|||

This report presents the counts (the frequencies), percentages, and a rough bar graph of the data. Note that the bar graph is constructed so that each "|" is worth 2.5 percentage points.

Example 2 – Multinomial Test Example

Suppose we want to test whether the proportion for Nevada is 50% higher than that for Virginia. The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Frequency Tables window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Frequency Tables window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Frequency Tables**. The Frequency Tables procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Frequency Tables window, select the **Variables tab**. (This is the default.)
- Double-click in the **Discrete Variables** text box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**. "State" will appear in the Variables box.

4 Specify the report format.

- Click on the **Format tab**.
- In **Variable Names**, select **Labels**.
- In **Value Labels**, select **Value Labels**.

5 Specify the Multinomial test.

- Click on the **Multinomial tab**.
- Check the **Multinomial Test** box.
- In **Expected Values for Multinomial Test** box, enter **60, 40**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Multinomial Test Output

Frequency Distribution of State

State	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
Nevada	88	88	58.67	58.67	||||||||||||||||||||||
Virginia	62	150	41.33	100.00	||||||||||||||||

Multinomial Test of State

State	Count	Expected Count	Actual Percent	Expected Percent	Chi-Square Amount
Nevada	88	90.00	58.67	60.00	0.0444
Virginia	62	60.00	41.33	40.00	0.0667

Chi-Square = 0.1111 with df = 1 Probability Level = 0.738883

The first table is the standard frequency table. The second table presents the results of the multinomial test. Note that in this case, the test is not signification.

Count

The number of observations (rows) in which the variable has the value reported on this line. This is the value of O_i.

Expected Count

The number of observations (rows) that would be obtained if the hypothesized proportions were followed exactly. This is the value of E_i.

Actual Percent

The percent that this category is of the total.

Expected Percent

The percentage that this category would have if the hypothesized proportions were followed exactly.

Chi-Square Amount

The amount that this line contributes to the Chi-square statistic. This is equal to

$$CS_i = \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the actual count and E_i is the expected count of the ith category.

Chi-Square

This is the value of the chi-square test statistic.

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

df

The degrees of freedom of the above test statistic. This is equal to the number of categories minus one.

Probability Level

This is the significance level of the multinomial test. If you are testing at an alpha of 0.05, you would reject the null hypothesis that the hypothesized proportions are true if this value is less than 0.05.

Example 3 – Tables of Counts and Percentages

This example will show how to obtain some of the other table formats that are available from this procedure. The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Frequency Tables window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Frequency Tables window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Frequency Tables**. The Frequency Tables procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Frequency Tables window, select the **Variables tab**. (This is the default.)
- Double-click in the **Discrete Variables** text box. This will bring up the variable selection window.
- Select **Garage, Fireplace,** and **Brick** from the list of variables and then click **Ok**. "Garage-Fireplace,Brick" will appear in the Discrete Variables box.

4 Specify the report format.

- Click on the **Format tab**.
- In **Variable Names**, select **Labels**.
- In **Value Label**s, select **Value Labels**.
- In **Label Justification**, select **Right**.
- In **Data Justification**, select **Right**.
- In **Tabs - First**, enter **2.0**.
- In **Tabs - Increment**, enter **0.75**.

5 Specify the reports.

- Click on the **Reports tab**.
- **Remove the check** from the **Frequency Tables Report**.
- Check **Row Percentage Report**.
- Check **Column Percentage Report**.
- Check **Row Percentage Report**.
- Check **Table Percentage Report**.
- Check **Column Percentage Plot**.

- Check **Row Percentage Plot**.
- Check **Table Percentage Plot**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Tables of Counts and Percentages Output

Row Percentages Section

	Values					
Variables	**0**	**0.5**	**1**	**2**	**3**	**Total**
Garage Size	4.7	0.0	65.3	28.7	1.3	100.0
Fireplaces	26.0	0.0	52.0	22.0	0.0	100.0
Brick Ratio	34.0	31.3	34.7	0.0	0.0	100.0
Total	21.6	10.4	50.7	16.9	0.4	100.0

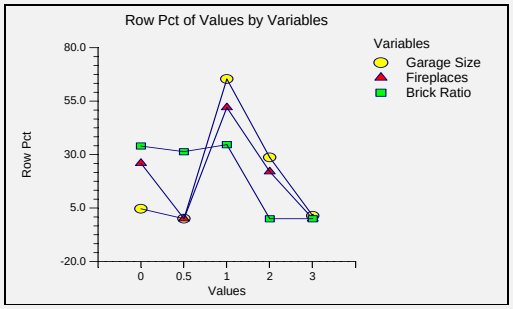
Column Percentages Section

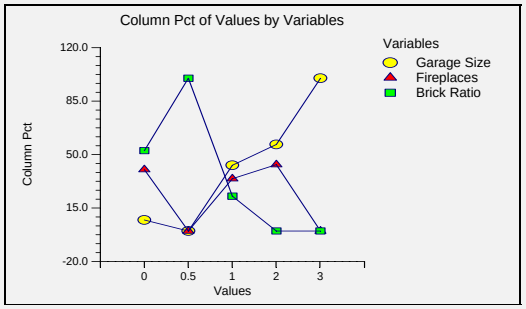
	Values					
Variables	**0**	**0.5**	**1**	**2**	**3**	**Total**
Garage Size	7.2	0.0	43.0	56.6	100.0	33.3
Fireplaces	40.2	0.0	34.2	43.4	0.0	33.3
Brick Ratio	52.6	100.0	22.8	0.0	0.0	33.3
Total	100.0	100.0	100.0	100.0	100.0	100.0

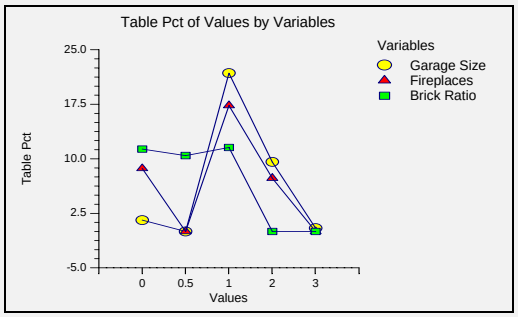
Table Percentages Section

	Values					
Variables	**0**	**0.5**	**1**	**2**	**3**	**Total**
Garage Size	1.6	0.0	21.8	9.6	0.4	33.3
Fireplaces	8.7	0.0	17.3	7.3	0.0	33.3
Brick Ratio	11.3	10.4	11.6	0.0	0.0	33.3
Total	21.6	10.4	50.7	16.9	0.4	100.0

Plot Section







This report presents tables containing various percentages for all variables selected. It also provides scatter plots of the percentage values.

Example 4 – Combined Tables

This example will show how to obtain a combined table of various counts and percentages. The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Frequency Tables window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Frequency Tables window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Frequency Tables**. The Frequency Tables procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Frequency Tables window, select the **Variables tab**. (This is the default.)
- Double-click in the **Discrete Variables** text box. This will bring up the variable selection window.
- Select **Garage, Fireplace,** and **Brick** from the list of variables and then click **Ok**. "Garage-Fireplace,Brick" will appear in the Discrete Variables box.

4 Specify the report format.

- Click on the **Format tab**.
- In **Variable Names,** select **Labels**.
- In **Value Labels**, select **Value Labels**.
- In **Label Justification**, select **Right**.
- In **Data Justification**, select **Right**.
- In **Tabs - First**, enter **2.0**.
- In **Tabs - Increment**, enter **0.75**.

5 Specify the reports.

- Click on the **Reports tab**.
- **Remove the check** from the **Frequency Tables Report**.
- Check **Show Combined Report**.
- Check **Counts**.
- Check **Row Percents**.
- Check **Column Percents**.
- Check **Table Percents**.
- Check **Double Space Report**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Combined Tables Output

Combined Table Section

Counts, Row Pct, Column Pct, Table Pct

	Values					
Variables	**0**	**0.5**	**1**	**2**	**3**	**Total**
Garage Size	7	0	98	43	2	150
	4.7	0.0	65.3	28.7	1.3	100.0
	7.2	0.0	43.0	56.6	100.0	33.3
	1.6	0.0	21.8	9.6	0.4	33.3
Fireplaces	39	0	78	33	0	150
	26.0	0.0	52.0	22.0	0.0	100.0
	40.2	0.0	34.2	43.4	0.0	33.3
	8.7	0.0	17.3	7.3	0.0	33.3
Brick Ratio	51	47	52	0	0	150
	34.0	31.3	34.7	0.0	0.0	100.0
	52.6	100.0	22.8	0.0	0.0	33.3
	11.3	10.4	11.6	0.0	0.0	33.3
Total	97	47	228	76	2	450
	21.6	10.4	50.7	16.9	0.4	100.0
	100.0	100.0	100.0	100.0	100.0	100.0
	21.6	10.4	50.7	16.9	0.4	100.0

This report presents a single table that combines the counts and percentages of all variables selected.

Chapter 501

Cross Tabulation

Introduction

This procedure produces tables of counts and percentages of the joint distribution of two variables that each have only a few distinct values. Such tables are known as contingency, cross-tabulation, or crosstab tables. When a breakdown of more than two variables is desired, you can specify up to eight break variables in addition to the two table variables. A separate table is generated for each unique set of values of these break variables.

This procedure serves as a summary reporting tool and is often used to analyze survey data. It also yields most of the popular contingency-table statistics such as chi-square, Fisher's exact, and McNemar's tests.

Types of Categorical Variables

Note that we will refer to two types of categorical variables: *By* and *Break***.** *Break* variables are used to split a database into subgroups. A separate table is generated for each unique set of values of the *Break* variables. The values of a *By* variable are used to define the rows and columns of the crosstab table. Two *By* variables are used per table, one defining the rows of the table and the other defining the columns.

Note that if you only want to use one *By* variable per table, you should use the *Frequency Table* procedure.

Data Structure

The data below are a subset of the *Real Estate Sales* database provided with the software. This (computer-simulated) data gives information including the selling price, the number of bedrooms, the total square footage (finished and unfinished), and the size of the lots for 150 residential properties sold during the last four months in two states. Only the first 8 of the 150 observations are displayed here.

RESALE dataset (subset)

State	Price	Bedrooms	TotalSqft	LotSize
Nev	260000	2	2042	10173
Nev	66900	3	1392	13069
Vir	127900	2	1792	7065
Nev	181900	3	2645	8484
Nev	262100	2	2613	8355

Missing Values

Missing values may be ignored or included in the table's counts, percentages, and statistical tests.

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tab

These two panels specify the variables that will be used in the analysis.

Specify at least one *Table Column* variable and at least one *Table Row* variable. The unique values of these two variables will form the columns and rows of the crosstab table. If more than one variable is specified in either section, a separate table will be generated for each combination of variables.

Four types of categorical variables may be specified:

1. Variables containing text values. These are called *Discrete Variables*.
2. Variables containing numeric values that are to be treated individually. For example, you might have used a set of index numbers like "1,2,3,4" to represent four states. These are also called *Discrete Variables.*
3. Variables containing numeric values that are to be grouped or combined into a set of predefined intervals. *You specify the interval boundaries*. For example, a variable containing age values might be grouped as "Under 21, 21 to 55, and Over 55." The key is that you specify the intervals. These are called *Numeric Variables (Limits).*
4. Variables containing numeric values that are to be combined into a set of computer-generated intervals. *You specify only the number of intervals*. The program determines a set of equal-length intervals based on the minimum and maximum found in the data. This format may cause problems since you do not set the interval boundaries directly. These are called *Numeric Variables (Width).*

Variables for Use in Table Columns and Rows

Discrete Variables

This option specifies those variables that contain text and numeric values that are to be treated as discrete variables (Types 1 or 2). Variables containing text values are always listed here. Variables containing numeric values are listed here if you want each unique value to be treated separately.

Numeric Variables (Width)

Use this option to specify variables that contain numeric values that are to be combined into a set of computer-generated intervals (Type 4). The intervals are specified in the three boxes: Number, Minimum, and Width. Note that you can specify one, two, or all three of these options.

Number

The number of intervals to be created. If not enough intervals are specified to reach the maximum data value, more intervals are added.

Minimum

The minimum value or the left boundary of the first interval. This value must be less than the minimum data value.

Width

This is the width of an interval. A data value X is in this interval if Lower Limit < X <= Upper Limit. If this is left blank, it is calculated from the Number, Minimum, and maximum data value.

Numeric Variables (Limits)

This specifies those variables that contain numeric values that are to be combined into a set of user-specified intervals (Type 3). The interval boundaries are specified as a list in the Interval Upper Limits box.

Interval Upper Limits

Specify the upper limits of the intervals, separated by commas. For example, you would enter "1,3,5" to specify the four intervals: Under 1, 1 to 3, 3 to 5, and Over 5.

The logic structure of the interval is:

Lower Bound < Value <= Upper Bound.

Note that a "1" would be included in the "Under 1" interval, not the "1 to 3" interval. Also, a "5" would be included in the "3 to 5" interval, not the "Over 5" interval.

Frequency Variable

Frequency Variable

This optional variable specifies the number of observations that each row represents. When omitted, each row represents a single observation. If your data is the result of previous summarization, you may want certain rows to represent several observations. Note that negative values are treated as a zero frequency and are omitted. Fractional values may be used. You may also think of this as a weighting variable.

Breaks Tab

This panel lets you specify up to eight break variables.

Select Break (Grouping) Variables

Break Variables

Specify one or more categorical variables whose distinct values will cause separate reports to be generated. Note that a separate set of reports (tables and plots) is generated for each unique set of values of these variables. Do not confuse these variables with the *Table Column* and *Table Row* variables, which specify the variables whose values will appear along the rows or columns of a particular table.

Missing Tab

This panel lets you specify up to five missing values (besides the default of blank). For example, '0', '9', or 'NA' may be missing values in your database.

Missing Value Options

Missing Values

Specify up to five missing values here.

- **Delete All**

  Indicates that you want the missing values totally ignored.

- **Include in Counts**

  Indicates that you want the number of missing values displayed, but you do not want them to influence any of the percentages.

- **Include in All**

  Indicates that you want the missing values treated just like any other category. They will be included in all percentages and counts.

Format Tab

The following options control the format of the reports.

Format Options

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want the table to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Label Justification

This option specifies whether the labels should be right or left justified above each column.

Data Justification

This option specifies whether the data should be right or left justified in each cell.

Split Column Headings

This option lets you select whether to split the column headings into two headings instead of one.

Tabs

These options let you specify the tab settings across the table. The output ruler is also modified by the settings of *Label Justification* and *Data Justification*.

First

Specifies the position of the first cell in inches. Note that the left-hand label always begins at 0.5 inches. Hence, the distance between this tab and 0.5 is the width provided for the row label information.

Maximum

Specifies the right border of the table. The number of tabs is determined based on *First*, the *Increment*, and this option. If you set this value too large, your table may not be printed correctly.

Increment

Specifies the width of a cell in inches.

Offset

The amount (inches) of offset to the right used with a decimal tab on a custom ruler so the data is aligned properly under the left-justified column labels.

Decimal Places

These options let you specify the number of decimal places used in the various items of the table.

Column-By

Specifies the number of decimal places displayed in the numeric *Table Columns* variable values. Note that *All* displays a single-precision (seven place accuracy).

Row-By

Specifies the number of decimal places displayed in the numeric *Table Rows* variable values. Note that *All* displays a single-precision (seven place accuracy).

Counts ... Std Resid

Specifies the number of decimal places displayed in each statistic. Note that *All* displays the default amount.

Reports Tab

These options control which of the available statistics are displayed.

Select Reports

List Report

Specify whether to display the List Report.

Fisher's Exact Test

Specify whether to display Fisher's exact test (2-by-2 tables only).

Armitage Proportion Trend Test

Specify whether to display Armitage proportion trend test (2-by-*k* tables only).

Select Statistics to be Displayed in Reports and Plots

Counts ... Std. Residual

For each of these statistics, you specify whether you want a numeric report, a plot, or both.

Chi-Square Stats

Specify whether to display the chi-square test of independence and related statistical tests.

Combined Report

These options specify the combined report of counts and percentages.

Show Combined Report

Specify whether to display this report.

Combined Report – Items on Report

Counts ... Std. Residuals

Specify whether to display these items on the combined report.

Combined Report – Options

Double Space Report

This option adds a blank row after each set of percentages in combined tables.

Plot Options Tab

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Connect Line(s)

Specifies whether connect the points with lines for easier interpretation of trends.

Plot Settings – Legend

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the appropriate default value.

Titles

Plot Title

This is the text of the title. The characters *{Y}*, *{X}*, and *{G}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Show Break as Title

Specifies whether the current values of any *Break* variables should be displayed as a second title line in the plot.

Symbols Tab

Specify the symbols used for each of the groups on the plots.

Plotting Symbols

Group 1-15

Specify the symbol used to designate a particular group. Double-click on a symbol or click on the button to the right of a symbol to specify the symbol's size, type, and color.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Standard Cross Tabulation Table

The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Cross Tabulation window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Cross Tabulation window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Cross Tabulation**. The Cross Tabulation procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Cross Tabulation window, select the **Variables tab**. (This is the default.)
- Double-click in the **Table Columns - Discrete Variables** text box. This will bring up the variable selection window.
- Click **Clear** and click **Ok**. The Table Columns - Discrete Variables box will be empty.
- Double-click in the **Table Rows - Discrete Variables** text box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**. "State" will appear in the Table Rows - Discrete Variables box.
- Double-click in the **Table Columns - Numeric Variables (Limits)** text box. This will bring up the variable selection window.
- Select **Price** from the list of variables and then click **Ok**. "Price" will appear in the Table Columns - Numeric Variables (Limits) box.
- In the **Table Columns - Interval Upper Limits** box, enter **100000, 200000, 300000**.

4 Specify the report format.

- Click on the **Format tab**.
- In **Variable Names**, select **Labels**.
- In **Value Labels**, select **Value Labels**.
- In **Label Justification**, select **Right**.
- In **Data Justification**, select **Right**.
- Check **Split Column Headings**.

- In the **Tabs - First** box, select **2.0**.
- In the **Decimal Places - Row %s**, select **1**.

5 Specify the reports.

- Click on the **Reports tab**.
- In **Row Percents**, select **Both**.
- In **Chi-Sqr Stats**, select **Omit**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

The following reports and charts will be displayed in the Output window.

Basic Cross Tabulation Output

Counts Section

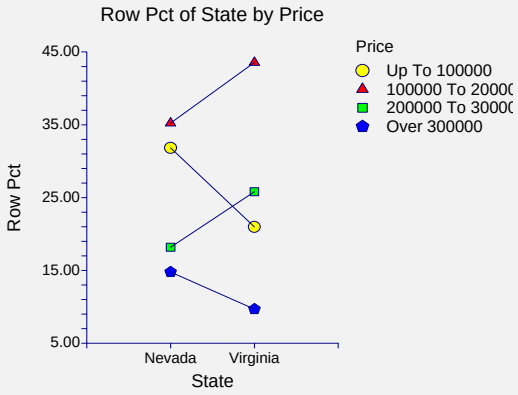
	Sales Price				
	Up to	100000	200000	Over	
State	100000	To 200000	To 300000	300000	Total
Nevada	28	31	16	13	88
Virginia	13	27	16	6	62
Total	41	58	32	19	150

The number of rows with at least one missing value is 0

Row Percentages Section

	Sales Price				
	Up to	100000	200000	Over	
State	100000	To 200000	To 300000	300000	Total
Nevada	31.8	35.2	18.2	14.8	100.0
Virginia	21.0	43.5	25.8	9.7	100.0
Total	27.3	38.7	21.3	12.7	100.0

The number of rows with at least one missing value is 0



This report presents tables of the various statistics. A plot of the row percentages is also displayed to make interpreting the row percentages easier. We will now define each of the possible statistics that may be displayed in these tables:

Count

The number of observations (rows) in the cell defined by the *By* variables. This is labelled O_{ij} in the formulas below.

Row Percent

The percent that this cell's count is of the row's total count. The total used depends on which missing value option was specified.

Column Percent

The percent that this cell's count is of the column's total count. The total used depends on which missing value option was specified.

Table Percent

The percent that this cell's count is of the total count of the table. The total used depends on which missing value option was specified.

Expected Value

The count that would be obtained if the hypothesis of row-column independence holds exactly. This is the value of E_{ij}.

$$E_{ij} = \frac{R_i C_j}{N}$$

where

$$R_i = \sum_j O_{ij}$$

$$C_j = \sum_i O_{ij}$$

$$N = \sum_{i,j} O_{ij}$$

Chi-Square

The amount that this cell contributes to the chi-square statistic. This and the next two items let you determine which cells impact the chi-square statistic the most.

$$CS_{ij} = \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Cell-Deviation

The observed count minus the expected count.

$$CD_{ij} = O_{ij} - E_{ij}$$

Std. Residual

The standardized residual is the *cell-deviation* scaled by the square root of the expected value:

$$SR_{ij} = \frac{\left(O_{ij} - E_{ij}\right)}{\sqrt{E_{ij}}}$$

Example 2 – Chi-Square Test and Related Statistics

The following example will demonstrate how to obtain a chi-square test for independence and related contingency table statistics. Our example generates a 2-by-2 table so that Fisher's exact test will also print. The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Cross Tabulation window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Cross Tabulation window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Cross Tabulation**. The Cross Tabulation procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Cross Tabulation window, select the **Variables tab**. (This is the default.)
- Double-click in the **Table Columns - Discrete Variables** text box. This will bring up the variable selection window.
- Click **Clear** and click **Ok**. The Table Columns - Discrete Variables box will be empty.
- Double-click in the **Table Rows - Discrete Variables** text box. This will bring up the variable selection window.
- Click **Clear** and click **Ok**. The Table Rows - Discrete Variables box will be empty.
- Double-click in the **Table Columns - Numeric Variables (Limits)** text box. This will bring up the variable selection window.
- Select **Price** from the list of variables and then click **Ok**. "Price" will appear in the Table Columns - Numeric Variables (Limits) box.
- In the **Table Columns - Interval Upper Limits** box, enter **150000**.
- Double-click in the **Table Rows - Numeric Variables (Limits)** text box. This will bring up the variable selection window.
- Select **TotalSqft** from the list of variables and then click **Ok**. "TotalSqft" will appear in the Table Rows - Numeric Variables (Limits) box.
- In the **Table Rows - Interval Upper Limits** box, enter **2000**.

4 Specify the report format.

- Click on the **Format tab**.
- In **Label Justification**, select **Right**.
- In **Data Justification**, select **Right**.
- Check **Split Column Headings**.

- In the **Tabs - First** box, select **2.0**.
- In the **Decimal Places - Row %s**, select **1**.

5 Specify the reports.

- Click on the **Reports tab**.
- In **Counts**, select **Omit**.
- In **Chi-Sqr Stats**, select **All**.
- Check **Fisher's Exact Test**.
- Check **Armitage Proportion Trend Test**.
- Check **Show Combined Report**.
- Check **Expected Values**.
- Check **Chi-Squares**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Combined Report and Contingency Statistics

Combined Report

Counts, Expected, Chi-Square

TotalSqft	Price Under 150000	Over 150000	Total
Under 2000	44	43	87
	40.0	47.0	87.0
	0.40	0.34	0.00
Over 2000	25	38	63
	29.0	34.0	63.0
	0.55	0.47	0.00
Total	69	81	150
	69.0	81.0	150.0
	0.95	0.81	1.76

The number of rows with at least one missing value is 0

Chi-Square Statistics Section

Chi-Square	1.745203	
Degrees of Freedom	1.000000	
Probability Level	0.186481	Accept Ho
Phi	0.107864	
Cramer's V	0.107864	
Pearson's Contingency Coefficient	0.107242	
Tschuprow's T	0.107864	
Lambda A .. Rows dependent	0.000000	
Lambda B .. Columns dependent	0.014493	
Symmetric Lambda	0.007576	
Kendall's tau-B	0.053423	
Kendall's tau-B (with correction for ties)	0.107864	
Kendall's tau-C	0.106133	
Gamma	0.217328	
Kappa reliability test	0.104792	
Kappa's standard error	0.079324	
Kappa's t value	1.321061	
McNemar's Test Statistic	4.764706	
McNemar's Degrees of Freedom	1.000000	
McNemar's Probability Level	0.029049	

Armitage Test for Trend in Proportions		
Ho: p1 = p2 = p3 = ... = pk		
Armitage S	-597	
Standard Error of S	453.4233	
Z-Value (Standardized S)	-1.316651	
Prob (Ha: Increasing Trend)	0.906022	Accept Ho
Prob (Ha: Decreasing Trend)	0.093978	Accept Ho
Prob (Ha: Any Trend)	0.187956	Accept Ho

Fisher's Exact Test Section

	P1	P2
Proportions	0.637681	0.530864
Difference (D0 = P1-P2)		0.106817
Correlation Coefficient		0.107864

Hypothesis	Prob Level	Test Type	Calculation Method
Ho: P1=P2			D=P1-P2 for a table
Ha: P1<P2	0.931755	One-Tailed	Sum of prob's of tables where D<=D0
Ha: P1>P2	0.123935	One-Tailed	Sum of prob's of tables where D>=D0
Ha: P1<>P2	0.245266	Two-Tailed	Sum of prob's of tables where |D|>=|D0|

Combined Report

The combined report simply includes all the information in one table rather than creating a separate table for each item selected.

Chi-Square Statistics Section

This section presents various contingency table statistics for studying the independence between rows and columns of the crosstab (contingency) table.

Chi-Square

The chi-square statistic is used to test independence between the row and column variables. Independence means that knowing the value of the row variable does not change the probabilities of the column variable (and vice versa). Another way of looking at independence is to say that the row percentages (or column percentages) remain constant from row to row (or column to column).

Note that this test requires large sample sizes to be accurate. An often quoted rule of thumb regarding sample size is that *none of the expected cell values can be less than five*. Although some users ignore the sample size requirement, you should be very skeptical of the test if you have cells in your table with zero counts. In the 2-by-2 case, consider using *Fisher's Exact Test* for small samples.

The formula for the chi-square test statistic is:

$$\chi^2_{df} = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Degrees of Freedom

This is the degrees of freedom, df, of the above chi-square test statistic. The formula is:

$$df = (r-1)(c-1)$$

where r is the number of rows and c is the number of columns in the table.

Probability Level

The probability of obtaining the above chi-square statistic or larger when the variables are independent. If you are testing at the alpha = 0.05 level of significance, this number must be less than 0.05 in order for the chi-square value to be significant. Significance means that the variables forming the rows and columns of the table are not independent.

Phi

A measure of association independent of the sample size. *Phi* ranges between 0 (no relationship) and 1 (perfect relationship). *Phi* was designed for two-by-two tables only. For larger tables, it has no upper limit and Cramer's V should be used instead. The formula is

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Cramer's V

A measure of association independent of sample size. This statistic is a modification of the Phi statistic so that it is appropriate for larger than two-by-two tables. V ranges between 0 (no relationship) and 1 (perfect relationship).

$$V = \sqrt{\frac{\phi^2}{K-1}}$$

where K is the lesser of the number of rows and the number of columns.

Pearson's Contingency Coefficient

A measure of association independent of sample size. It ranges between 0 (no relationship) and 1 (perfect relationship). For any particular table, the maximum possible depends on the size of the table (a 2-by-2 table has a maximum of 0.707), so it should only be used to compare tables with the same dimensions. The formula is

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Tschuprow's T

A measure of association independent of sample size. This statistic is a modification of the Phi statistic so that it is appropriate for larger than two-by-two tables. T ranges between 0 (no relationship) and 1 (perfect relationship), but 1 is only attainable for square tables. The formula is

$$T = \sqrt{\frac{\phi^2}{\sqrt{(r-1)(c-1)}}}$$

Lambda A - Rows dependent

This is a measure of association for cross tabulations of nominal-level variables. It measures the percentage improvement in predictability of the dependent variable (row variable or column variable), given the value of the other variable (column variable or row variable). The formula is

$$\lambda_a = \frac{\sum_i \max(O_{ij}) - \max(R_i)}{N - \max(R_i)}$$

Lambda B - Columns dependent

See Lambda A above. The formula is

$$\lambda_b = \frac{\sum_j \max(O_{ij}) - \max(C_j)}{N - \max(C_j)}$$

Symmetric Lambda

This is a weighted average of the *Lambda A* and *Lambda B* above. The formula is

$$\lambda = \frac{\sum_i \max(O_{ij}) + \sum_j \max(O_{ij}) - \max(R_i) - \max(C_j)}{2N - \max(R_i) - \max(C_j)}$$

Kendall's tau-B

This is a measure of correlation between two ordinal-level (rankable) variables. It is most appropriate for square tables. To compute this statistic, you first compute two values, P and Q, which represent the number of concordant and discordant pairs, respectively. The formula is

$$\tau_b = \frac{P - Q}{N(N-1)/2}$$

Kendall's tau-B (with correction for ties)

This is the same as the above, except a correction is made for the case when ties are found in the data.

Kendall's tau-C

This is used in the case where the number of rows does not match the number of columns. The formula is

$$\tau_c = \frac{P - Q}{N^2(k-1)/(2k)}$$

where *k* is the minimum of *r* and *c*.

Gamma

This is another measure based on concordant and discordant pairs. The formula is

$$\gamma = \frac{P - Q}{P + Q}$$

Kappa Reliability Test

Kappa is a measure of association (correlation or reliability) between two measurements of the same individual when the measurements are categorical. It tests if the counts along the diagonal are significantly large. Because Kappa is used when the same variable is measured twice, it is only appropriate for a square tables. Kappa is often used to study the agreement of two raters such as judges or doctors. Each rater classifies each individual into one of *k* categories.

Rules-of-thumb for kappa: values less than 0.40 indicate low association; values between 0.40 and 0.75 indicate medium association; and values greater than 0.75 indicate high association between the two raters. The formulas for Kappa, its standard error, and associated t-value are

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$s_\kappa = \frac{1}{(1 - p_e)} \sqrt{\frac{p_e + p_e^2 - \sum_i p_{i\cdot} p_{\cdot i} (p_{\cdot i} + p_{i\cdot})}{N}}$$

$$t_\kappa = \frac{\kappa}{s_\kappa}$$

where

$$p_o = \frac{\sum_i O_{ii}}{N}$$

$$p_e = \frac{\sum_i R_i C_i}{N^2}$$

$$p_{\cdot j} = \frac{\sum_j O_{ij}}{N}$$

$$p_{i\cdot} = \frac{\sum_j Oij}{N}$$

McNemar Test

The McNemar test was first used to compare two proportions that are based on matched samples. Matched samples occur when individuals (or matched pairs) are given two different treatments, asked two different questions, or measured in the same way at two different points in time. Match pairs can be obtained by matching individuals on several other variables, by selecting two people from the same family (especially twins), or by dividing a piece of material in half.

The McNemar test has been extended so that the measured variable can have more than two possible outcomes. It is then called the *McNemar test of symmetry*. It tests for symmetry around the diagonal of the table. The diagonal elements of the table are ignored.

The test is computed for square tables only. The formula is

$$\chi^2_{MC} = \frac{1}{2} \sum_i \sum_j \frac{(O_{ij} - O_{ji})^2}{(O_{ij} + O_{ji})}$$

The degrees of freedom of this chi-square statistic is given by $r(r\text{-}1)/2$. The probability level gives the significance level. That is, reject the hypothesis of symmetry about the diagonal if the reported probability level is less than some predetermined level, such as 0.05.

Armitage Test for Trend in Proportions

This is a test for linear trend in proportions proposed by Armitage (1955). The test may be used when you have exactly two rows or two columns in your table. This procedure tests the hypothesis that there is a linear trend in the proportion of successes. That is, the true proportion of successes increases (or decreases) as you move from row to row (or column to column).

The test statistic, *S*, is standardized to a normal z-value by dividing by the estimated standard error of *S* (which we label *V* below). This z-value can be tested using the standard-normal distribution.

When there are two columns and we want to test for the presence of a trend in proportions down the rows, the calculations for this test are as follows:

$$z = \frac{S}{\sqrt{V}}$$

where

$$S = A - B$$

$$A = \sum_{j=1}^{r-1} O_{j2} \sum_{i=j+1}^{r} O_{i1}$$

$$B = \sum_{j=1}^{r-1} O_{j1} \sum_{i=j+1}^{r} O_{i2}$$

$$V = \frac{C_1 C_2 \left(N^3 - \sum_{i=1}^{r} R_i^3 \right)}{3N(N-1)}$$

Fisher's Exact Test

This test was designed to test the hypothesis that the two column percentages in a 2-by-2 table are equal. It is especially useful when sample sizes are small (even zero in some cells) and the chi-square test is not appropriate.

Exact probability levels are given for one-sided and two-sided alternatives. You would reject the null hypothesis of equality of proportions when the reported probability level is less than a stated level, such as 0.05.

The calculation of these probability levels is made by calculating how many of the possible tables that may be constructed from the marginal totals given in this table support the alternative hypothesis.

Example 3 – List Report

The list format was designed for situations in which you want to transfer a summarized table to another program. This format creates a vertical listing of the counts in a format that is easy to copy and paste into another ***NCSS*** database or into other programs. The data for this example are found in the RESALE database. You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Cross Tabulation window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Cross Tabulation window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Cross Tabulation**. The Cross Tabulation procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Cross Tabulation window, select the **Variables tab**.
- Double-click in the **Table Columns - Discrete Variables** text box. This will bring up the variable selection window.
- Select **City** from the list of variables and then click **Ok**. "City" will appear in the Table Columns - Discrete Variables box.
- Double-click in the **Table Rows - Discrete Variables** text box. This will bring up the variable selection window.
- Select **Brick** from the list of variables and then click **Ok**. "Brick" will appear in the Table Columns - Discrete Variables box.

4 Specify a break variable.

- On the Cross Tabulation window, select the **Breaks tab**.
- Double-click in the **first Break Variables** text box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**. "State" will appear in the text box.

5 Specify the reports.

- Click on the **Reports tab**.
- In **Counts**, select **Omit**.
- In **Chi-Sqr Stats**, select **Omit**.
- Check **List Report**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

List Report

List Report

State	City	Brick	Count
Nev	1	0	7
Nev	1	0.5	9
Nev	1	1	11
Nev	2	0	16
Nev	2	0.5	19
Nev	2	1	14
Nev	3	0	5
Nev	3	0.5	2
Nev	3	1	5
Vir	4	0	11

(report continues)

List Report

This report gives the count (frequency) for each unique combination of the *By* and *Break* variables, taken together.

Example 4 – Cross Tabs on Summarized Data

This example will demonstrate how to enter and analyze a contingency table that has already been summarized. Suppose the following data from a study of the impact of three drugs on a certain disease are available.

		Drug Type	
Disease	**Type 1**	**Type 2**	**Type 3**
Yes	15	28	44
No	4	7	9

These data were entered in an ***NCSS*** database called DrugStudy.s0 in three variables. The database appears as follows:

DRUGSTUDY dataset

DrugType	Disease	Count
1	Yes	15
2	Yes	28
3	Yes	44
1	No	4
2	No	7
3	No	9

Notice that we have used three variables: one containing the column identification (DrugType), one containing the row identification (Disease), and one containing the counts or frequencies (Count).

You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Cross Tabulation window.

1 Open the DRUGSTUDY dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **DrugStudy.s0**.
- Click **Open**.

2 Open the Cross Tabulation window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Cross Tabulation**. The Cross Tabulation procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Cross Tabulation window, select the **Variables tab**. (This is the default.)
- Double-click in the **Table Columns - Discrete Variables** text box. This will bring up the variable selection window.
- Select **DrugType** and click **Ok**. The Table Columns - Discrete Variables box will contain "DrugType".
- Double-click in the **Table Rows - Discrete Variables** text box. This will bring up the variable selection window.
- Select **Disease** and click **Ok**. The Table Rows - Discrete Variables box will contain "Disease".
- Double-click in the **Frequency Variable** text box. This will bring up the variable selection window.
- Select **Count** and click **Ok**. The Frequency Variable box will contain "Count".

4 Specify the reports.

- Click on the **Reports tab**.
- In **Counts**, select **Omit**.
- In **Chi-Sqr Stats**, select **Chi-Square**.
- Check **Show Combined Report**.
- Check **Counts**.
- Check **Row Percents**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Cross Tabulation Output

Combined Report
Counts, Row Pct

	DrugType			
Disease	**1**	**2**	**3**	**Total**
No	4	7	9	20
	20.0	35.0	45.0	100.0
Yes	15	28	44	87
	17.2	32.2	50.6	100.0
Total	19	35	53	107
	17.8	32.7	49.5	100.0

The number of rows with at least one missing value is 0

Chi-Square Statistics Section

Chi-Square	0.211145	
Degrees of Freedom	2.000000	
Probability Level	0.899809	Accept Ho

WARNING: At less one cell had an expected value less than 5.

Combined Report

We arbitrarily selected the Combined Report and Chi-Square Statistics. Of course, you could select any of the reports you wanted to see.

Chapter 505

Item Analysis

Introduction

This procedure performs item analysis. Item analysis studies the internal reliability of a particular instrument (test, survey, questionnaire, etc.). This instrument usually consists of several questions (items) which are answered by a group of respondents. Issues that arise include whether the instrument measures what was intended (does a particular IQ test reliably measure an individual's intelligence?), whether it produces the same results when it is administered repeatedly, whether it contains cultural biases, and so on.

Item analysis is not the same as item response analysis. Item response analysis is concerned with the analysis of questions on a test which can be scored as either right or wrong. The Item Response Analysis program, discussed elsewhere, conducts this type of analysis.

Discussion

Because of the central role of measurement in science, scientists of all disciplines are concerned with the accuracy of their measurements. Item analysis is a methodology for assessing the accuracy of measurements that are obtained in the social sciences where precise measurements are often difficult to secure. The accuracy of a measurement may be divided into two dimensions: validity and reliability. The *validity* of an instrument refers to whether it accurately measures the attribute of interest. The *reliability* of an instrument concerns whether it produces identical results in repeated applications. An instrument may be reliable but not valid. However, it cannot be valid without being reliable.

The methods described here assess the reliability of an instrument. They do not assess its validity. This should be kept in mind when using the techniques of item analysis since they address reliability, not validity.

An instrument may be valid for one attribute but not for another. For example, a driver's license exam may accurately measure an individual's ability to drive. However, it does not accurately measure that individual's ability to do well in college. Hence the exam is reliable and valid for measuring driving ability. It is reliable and invalid for measuring success in college.

Several methods have been proposed for assessing the reliability of an instrument. These include the retest method, alternative-form method, split-halves method, and the internal consistency method. We will focus on internal consistency here.

Cronbach's alpha (or coefficient alpha) is the most popular of the internal consistency coefficients. It is calculated as follows:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K}\sigma_{ii}}{\sum_{i=1}^{K}\sum_{j=1}^{K}\sigma_{ij}}\right]$$

where K is the number of items (questions) and σ_{ij} is the estimated covariance between items i and j. Note the σ_{ii} is the variance (not standard deviation) of item i.

If the data are standardized by subtracting the item means and dividing by the item standard deviations before the above formula is used, we get the standardized version of Cronbach's alpha. A little algebra will show that this is equivalent to the following calculations based directly on the correlation matrix of the items:

$$\alpha = \frac{K\overline{\rho}}{1+\overline{\rho}(K-1)}$$

where $\overline{\rho}$ is the average of all the correlations among the K items.

Cronbach's alpha has at least three interpretations.

1. Cronbach's alpha is equal to the average value of alpha coefficients obtained for all possible combinations of dividing $2K$ items into two groups of K items each and calculating the two-half tests.
2. Cronbach's alpha estimates the expected correlation of one instrument with an alternative form containing the same number of items.
3. Cronbach's alpha estimates the expected correlation between an actual test and a hypothetical test which may never by written.

Since Cronbach's alpha is suppose to be a correlation, it should range between -1 and 1. However, it is possible for alpha to be less than -1 when several of the covariances are relatively large, negative numbers. In most cases, alpha is positive, although negative values arise occasionally. What value of alpha should be achieved? Carmines (1990) stipulates that as a rule, a value of at least 0.8 should be achieved for widely used instruments. An instrument's alpha value may be improved by either adding more items or by increasing the average correlation among the items.

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown in the table below These data are contained in the ITEM database. These data represent the responses of sixteen individuals to a four-item questionnaire.

ITEM dataset

Item1	Item2	Item3	Item4
1	3	2	1
2	2	2	3
1	3	2	2
3	3	3	3
1	1	2	2
3	3	3	1
2	2	1	2
1	1	2	1
1	3	1	2
1	1	2	2
5	3	2	2
1	1	2	1
1	3	2	2
1	3	3	1
1	3	2	1
1	3	1	1

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Item Variables

Item Variables

Specify two or more variables to be analyzed. These are the variables containing each individual's responses to the questionnaire. Each variable represents an item (question). Each row represents an individual.

Frequency Variable

Frequency Variable

This optional variable contains the frequency (count) to be assigned to this row. Normally, each row receives a count of one. The values in this variable replace this default value.

Options

Zero

Specify the value used as zero by the numerical routines. Because of round-off problems, values less than this amount (in absolute value) are changed to zero during the calculations.

Reports Tab

The following options control the format of the reports that are displayed.

Select Reports

Reliability Report ... Covariance Report

Indicate whether to display the indicated report.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Item Analysis

This section presents an example of how to run an analysis of the data contained in the ITEM database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Item Analysis window.

1 Open the ITEM dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Item.s0**.
- Click **Open**.

2 Open the Item Analysis window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Item Analysis**. The Item Analysis procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Item Analysis window, select the **Variables tab**.
- Double-click in the **Item Variables** box. This will bring up the variable selection window.
- Select **Item1** to **Item4** from the list of variables and then click **Ok**. "Item1-Item4" will appear in the Item Variables box.

4 Specify the reports.

- On the Item Analysis window, select the **Reports tab**.
- Check all reports options so that all of the reports will be displayed.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Reliability Section

Reliability Section

	Item Values		If This Item is Omitted				R2
Variable	Mean	Standard Deviation	Total Mean	Total Std.Dev.	Coef Alpha	Corr Total	Other Items
Item1	1.625	1.147461	6.0625	1.340087	0.0974	0.4932	0.2556
Item2	2.375	0.8850612	5.3125	1.778342	0.4506	0.2171	0.0921
Item3	2	0.6324555	5.6875	1.922455	0.4464	0.2193	0.0869
Item4	1.6875	0.7041543	6	1.897367	0.4583	0.1996	0.1371
Total			7.6875	2.15155	0.4704		

Cronbacks Alpha 0.470447 Std. Cronbachs Alpha 0.444639

This report shows important features of the reliability of the items on the instrument.

Mean
The item average.

Standard Deviation
The item standard deviation with divisor (n-1).

Total Mean
The average total of the other items when this item is ignored.

Total Std.Dev.
The standard deviation of the total of the other items when this item is ignored.

Coef Alpha
This is the value of Cronbach's alpha when this item is omitted.

Corr Total
This is the correlation between this item and the total of all other items. If this correlation is high, say greater than 0.95, then this item is redundant and might be omitted.

R2 Other Items
This is the R-Squared that results if this item is regressed on the other items. If this value is high, say greater than 0.95, then this item is redundant and might be omitted.

Cronbach's Alpha
Cronbach's alpha (or *coefficient alpha*) is a measure of internal reliability. Since Cronbach's alpha is a correlation, it can range between -1 and 1. In most cases it is positive, although negative values arise occasionally.

What value of alpha should be achieved? Carmines (1990) stipulates that as a rule, a value of at least 0.8 should be achieved for widely used instruments. An instrument's alpha value may be improved by either adding more items or by increasing the average correlation among the items.

Std. Cronbach's Alpha
If the data are standardized by subtracting the item means and dividing by the item standard deviations before the above formula is used, we obtain the standardized version of Cronbach's alpha.

Count Distribution Section

Count Distribution Section

Variable	1	2	3	5
Item1	11	2	2	1
Item2	4	2	10	0
Item3	3	10	3	0
Item4	7	7	2	0
Total	25	21	17	1

This report shows the number of times each response was chosen for each item.

Percentage Distribution Section

Percentage Distribution Section

Variable	1	2	3	5
Item1	68.75	12.50	12.50	6.25
Item2	25.00	12.50	62.50	0.00
Item3	18.75	62.50	18.75	0.00
Item4	43.75	43.75	12.50	0.00
Total	39.06	32.81	26.56	1.56

This report shows the percentages of each of the possible responses for each item.

Item Detail Section

Item Detail Section for Item1

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1	11	68.75	68.75	|||||||||||||||||||||||||||||||||||||||||||||
2	2	12.50	81.25	|||||||||
3	2	12.50	93.75	|||||||||
5	1	6.25	100.00	|||||
Total	16			

Item Detail Section for Item2

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1	4	25.00	25.00	|||||||||||||||||
2	2	12.50	37.50	|||||||||
3	10	62.50	100.00	|||||||||||||||||||||||||||||||||||||||||
5	0	0.00	100.00		
Total	16			

Item Detail Section for Item3

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1	3	18.75	18.75	|||||||||||||
2	10	62.50	81.25	|||||||||||||||||||||||||||||||||||||||||
3	3	18.75	100.00	|||||||||||||
5	0	0.00	100.00		
Total	16			

Item Detail Section for Item4

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1	7	43.75	43.75	|||||||||||||||||||||||||||||
2	7	43.75	87.50	|||||||||||||||||||||||||||||
3	2	12.50	100.00	|||||||||
5	0	0.00	100.00		
Total	16			

This report provides an individual break down of the responses to each item.

Correlation Section

Correlation Section

	Item1	Item2	Item3	Item4
Item1	1.000000	0.278989	0.275589	0.340351
Item2	0.278989	1.000000	0.119098	-0.013371
Item3	0.275589	0.119098	1.000000	0.000000
Item4	0.340351	-0.013371	0.000000	1.000000

Cronbacks Alpha 0.470447 Std. Cronbachs Alpha 0.444639

This report presents the correlations between each pair of items.

Covariance Section

Covariance Section

	Item1	Item2	Item3	Item4
Item1	1.316667	0.2833333	0.2	0.275
Item2	0.2833333	0.7833334	6.666667E-02	-8.333334E-03
Item3	0.2	6.666667E-02	0.4	0
Item4	0.275	-8.333334E-03	0	0.4958333

Cronbacks Alpha 0.470447 Std. Cronbachs Alpha 0.444639

This report presents the covariances between each pair of items.

Chapter 506

Item Response Analysis

Introduction

This procedure performs item response analysis. Item response analysis is concerned with the analysis of questions on a test which can be scored as either right or wrong.

Item Response Analysis is not the same as item analysis. Item analysis studies the internal reliability of a particular instrument (test, survey, questionnaire, etc.). The Item Analysis program, discussed elsewhere, conducts this type of analysis.

Discussion

Let A_j represent the j[th] individual's ability to perform a certain task. This ability may represent intelligence, math aptitude, geography knowledge, etc. Define the *logistic item characteristic curve* (ICC) as follows:

$$P_i(A_j) = \frac{1}{1 + e^{-Z_{ij}}}$$

where

$$Z_{ij} = d_i A_j + a_i$$

Note that $P_i(A_j)$ is the probability that individual j with ability A_j marks item i correctly. Z_{ij} is called the logit. An items difficulty may be calculated using:

$$b_i = \frac{-d_i}{a_i}$$

This model is similar to the usual logistic regression model. A new problem has arisen in that the ability values, the A_j, are unknown and must be estimated. The program uses the Bock-Aikin (1981) MMLE/EM algorithm as provided in Baker (1992).

Data Structure

The data are entered in two or more variables. Only numeric values are allowed. Also, a variable containing the correct answers must be entered. An optional variable containing row labels may also be entered.

Note that the answers correspond to items by position number. Thus the answer to the item contained in variable one is in row one, the answer to the item contained in variable two is in row two, and so on.

An example of data appropriate for this procedure is shown in the table below. These data are contained in the ITEM database. These data represent the responses of sixteen individuals to a four-item test.

ITEM dataset

Item1	Item2	Item3	Item4	Answer	Name
1	3	2	1	1	Eric
2	2	2	3	3	Charlotte
1	3	2	2	2	Janet
3	3	3	3	1	Julie
1	1	2	2		Katie
3	3	3	1		Ryan
2	2	1	2		Tyson
1	1	2	1		Jordan
1	3	1	2		Bob
1	1	2	2		Linda
5	3	2	2		Mike
1	1	2	1		Michell
1	3	2	2		Randy
1	3	3	1		Jerry
1	3	2	1		Chris
1	3	1	1		Holly

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Data Variables

Item Variables

Specify the variables to be analyzed. These are the variables containing each individual's answers to the questions on the test. Each variable represents a question. Each row represents an individual.

Note that only numeric variables are allowed here.

Answer Variable

This variable contains a list of the correct answers. The answers in this column correspond to the questions by position. The value in row one is the correct response for variable one, the value in

row two is the correct response for variable two, and so on. If your questions start with variable five (variables one - four are identification variables), your correct answers must start in row five!

If you have already scored the test and you want to enter simply right or wrong, use zero for wrong, one for right, and put all ones in this variable.

Other Variables

Frequency Variable

This optional variable contains the frequency (count) to be assigned to this row. Normally, each row receives a count of one. The values in this variable replace this default value.

Label Variable

This optional variable contains row labels that are used to label various reports.

Options

Iterations

This option specifies the maximum number of iterations. Experience has shown that a value of 50 or more is necessary.

Zero

Specify the value used as zero by the numerical routines. Because of round-off error, values less than this amount (in absolute value) are changed to zero during the calculations.

Reports Tab

The following options control the reports and plots that are displayed.

Select Reports

Counts Report ... Abilities Report

Indicate whether to display the indicated reports.

Select Plots

Item Response Plot

Indicate whether to display the item response plot.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Plot Options

Smoothing Interval

This value is used in creating the points that appear on the IRC plots. If an individual's ability level falls within plus or minus this amount of the ability value, their response is included in the calculation of the percent correct.

Ability Data Points

The number of data points displayed along the ability scale. The logistic curve is trying to fit these values.

Plots Per Row

This option controls the size of the plots by specifying the number of plots to display across the page.

IRC Plot Tab

This panel specifies the item response curve (IRC) plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Specifies the plotting symbols used.

Titles

Plot Title

This is the text of the title. The characters {Y} are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Storage Tab

The estimated abilities may be stored on the current database for further analysis. The data are automatically stored while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Select Variables for Data Storage

Ability Variable

This option lets you designate the variable that should receive the estimated ability of each individual.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Item Response Analysis

This section presents an example of how to run an analysis of the data contained in the ITEM database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Item Response Analysis window.

1 Open the Item dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Item.s0**.
- Click **Open**.

2 Open the Item Response Analysis window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Item Response Analysis**. The Item Response Analysis procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Item Response Analysis window, select the **Variables tab**.
- Double-click in the **Item Variables** box. This will bring up the variable selection window.
- Select **Item1** to **Item4** from the list of variables and then click **Ok**. “Item1-Item4” will appear in the Item Variables box.
- Double-click in the **Answer Variable** box. This will bring up the variable selection window.
- Select **Answers** from the list of variables and then click **Ok**. “Answers” will appear in the Answer Variable box.
- Double-click in the **Label Variable** box. This will bring up the variable selection window.
- Select **Name** from the list of variables and then click **Ok**. “Name” will appear in the Label Variable box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Count Distribution Section

Count Distribution Section

Variable	1	2	3	5
Item1	11	2	2	1
Item2	4	2	10	0
Item3	3	10	3	0
Item4	7	7	2	0
Total	25	21	17	1

This report shows the number of times each response was chosen for each item.

Percentage Distribution Section

Percentage Distribution Section

Variable	1	2	3	5
Item1	68.75	12.50	12.50	6.25
Item2	25.00	12.50	62.50	0.00
Item3	18.75	62.50	18.75	0.00
Item4	43.75	43.75	12.50	0.00
Total	39.06	32.81	26.56	1.56

This report shows the percentages of each of the possible responses for each item.

Item Detail Section

Item Detail Section for Item1

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1 (Correct)	11	68.75	68.75	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2	2	12.50	81.25	|IIIIIIII
3	2	12.50	94.75	|IIIIIIII
5	1	6.25	100.00	|IIII
Total	16			

Item Detail Section for Item2

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1	4	25.00	25.00	|IIIIIIIIIIIIIIIII
2	2	12.50	37.50	|IIIIIIII
3 (Correct)	10	62.50	100.00	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
5	0	0.00	100.00	|
Total	16			

Item Detail Section for Item3

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1	3	18.75	18.75	|IIIIIIIIIIII
2 (Correct)	10	62.50	81.25	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3	3	18.75	100.00	|IIIIIIIIIIII
5	0	0.00	100.00	|
Total	16			

Item Detail Section for Item4

Value	Count	Individual Percent	Cumulative Percent	Percent Bar Chart
1 (Correct)	7	43.75	43.75	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII
2	7	43.75	87.50	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII
3	2	12.50	100.00	|IIIIIIII
5	0	0.00	100.00	|
Total	16			

This report provides an individual break down of the responses to each item.

Item Response Estimation Section

Item Response Estimation Section

Variable	Intercept	Discrimination Parameter (Slope)	Ability at which P(Correct)=0.5 (Difficulty)
Item1	5.977607	12.677396	-0.471517
Item2	0.510955	-0.025970	19.675165
Item3	0.577043	0.775644	-0.743954
Item4	-0.296144	0.878906	0.336946

This report gives the results of the IRT estimation for each item. The columns of the report are defined as follows.

Variable

The name of the item (question).

Intercept

The estimated intercept in the logistic ICC model. This is the value of a_i.

Discrimination Parameter (Slope)

The estimated slope in the logistic ICC model. This is the value of d_i. This value is sometimes referred to as the *Discrimination Parameter* of the item.

Ability at which P(Correct)=0.5 (Difficulty)

The estimated value of b_i. It is the ability level at which 50% of those responding are able to get this question right. It is sometimes called the *Difficulty* of the item.

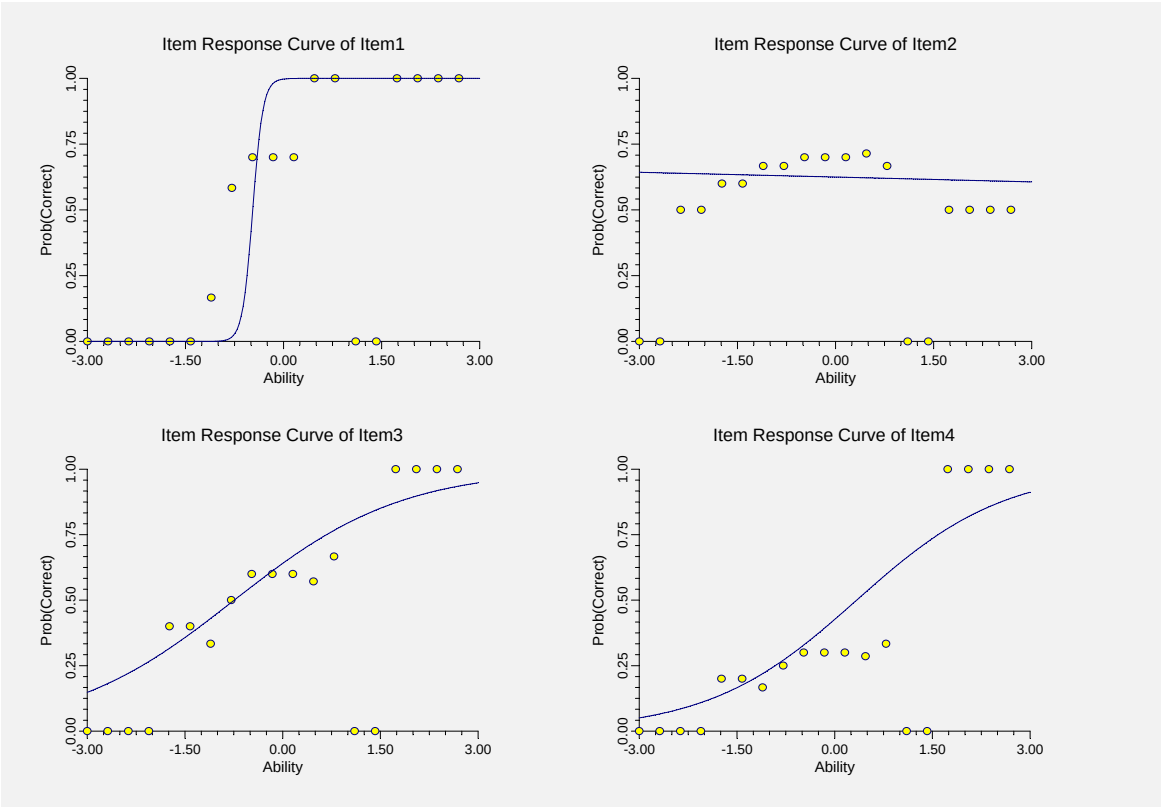
Estimated Abilities Section

Estimated Abilities Section

Row	Name	Number Correct	Percent Correct	Ability
1	Eric	4	100.00	2.500000
2	Charlotte	1	25.00	-0.800205
3	Janet	3	75.00	-0.077078
4	Julie	1	25.00	-1.552304
5	Katie	2	50.00	-0.053399
6	Ryan	2	50.00	-0.777427
7	Tyson	0	0.00	-1.536428
8	Jordan	3	75.00	2.500000
9	Bob	2	50.00	-0.260121
10	Linda	2	50.00	-0.053399
11	Mike	2	50.00	-0.809005
12	Michell	3	75.00	2.500000
13	Randy	3	75.00	-0.077078
14	Jerry	3	75.00	0.040460
15	Chris	4	100.00	2.500000
16	Holly	3	75.00	0.040460

This report gives each individual's score and estimated ability. Remember that the test score may not correlate exactly with ability since the ability rating depends not only on how many questions are answered correctly, but also on which questions are answered correctly.

IRC Plots Section



These plots show the logistic item characteristic curve for each question as a solid line. The plotted points show the proportion of individuals with ability in a small neighborhood of the plotted ability that got the question right.

Note that the vertical axis gives the probability that an individual answers the question correctly and the horizontal axis gives their ability.

In most cases, the plots will show the familiar S-curve shape that is exhibited here. In some cases, however, the plots may look different.

506-10 Item Response Analysis

Chapter 510

One Proportion

Introduction

This program computes confidence limits and hypothesis tests for a single proportion. For example, you might want confidence limits for the proportion of individuals with the common cold who took ascorbic acid (vitamin C) and recovered within twenty-four hours. You might want to test the hypothesis that more than 70% of a group of individuals with the common cold recovered immediately after taking ascorbic acid.

Exact results, based on the binomial distribution, are calculated. Approximate results, based on the normal approximation to the binomial distribution are also given. Of course, the exact results are preferable to the approximate results and should always be used. The approximate results are given because they are commonly presented in elementary statistical texts. They were developed in the pre-computer age when the calculations were carried out by hand (or on a hand calculator). Now that the exact results are available, there is no advantage in using the approximations. (It may be interesting to compare the results to see how close the approximations are).

The Binomial Model

Binomial data must exhibit the following four conditions:

1. The response can take on only one of two possible values. This is a binary response variable.
2. The response is observed a known number of times. Each replication is called a Bernoulli trial. The number of replications is labeled *n*. The number of responses out of the *n* total that exhibit the outcome of interest is labeled *X*. Thus *X* takes on the possible values 0, 1, 2, ..., n.
3. The probability that a particular outcome (a success) occurs is constant for each trial. This probability is labeled *p*.
4. The trials are independent. The outcome of one trial does not influence the outcome of the any other trial.

The binomial probability, *b(x; n, p)*, is calculated using:

$$b(x;n,p)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

where

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

The estimate of p from a sample is labeled $\hat{p}$ and is estimated using:

$$\hat{p} = \frac{X}{n}$$

In practice, p and $\hat{p}$ are used interchangeably.

Confidence Limits

Using a mathematical relationship (see Ostle(1988), page 110) between the F distribution and the cumulative binomial distribution, the lower and upper confidence limits of a 100 (1-α)% confidence interval are given by:

$$LCL = \frac{XF_{[\alpha/2],[2X,2(n-X+1)]}}{(n-X+1)+XF_{[\alpha/2],[2X,2(n-X+1)]}}$$

and

$$UCL = \frac{(X+1)F_{[1-\alpha/2],[2(X+1),2(n-X)]}}{(n-X)+(X+1)F_{[1-\alpha/2],[2(X+1),2(n-X)]}}$$

Note that although these limits are based on direct calculation of the binomial distribution, they are only 'exact' for a few values of alpha. Otherwise, these limits are conservative (wider than necessary).

These limits may be approximated using the normal approximation to the binomial as

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If a correction for continuity is added, the above formula becomes

$$\hat{p} \pm \left(z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2n} \right)$$

Although these two approximate confidence intervals are found in many elementary statistics books, they are not recommended. For example, Newcombe (1998) made a comparative study of seven confidence interval techniques and these methods came in last.

Instead, Newcombe (1998) recommended the Wilson Score confidence interval method because of its performance. The Wilson Score confidence interval is calculated using

$$\frac{\left(2n\hat{p} + z_{\alpha/2}^2\right) \pm z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\hat{p}\left(1-\hat{p}\right)}}{2\left(n + z_{\alpha/2}^2\right)}$$

Hypothesis Tests

Three sets of statistical hypotheses may be formulated:

1. H_0: $p = p_0$ versus H_A: $p \neq p_0$; this is often called the *two-tailed test*.
2. H_0: $p \leq p_0$ versus H_A: $p > p_0$; this is often called the *upper-tailed test*.
3. H_0: $p \geq p_0$ versus H_A: $p < p_0$; this is often called the *lower-tailed test*.

The exact p-values for each of these situations may be computed as follows:

1. $P\left(|\tilde{p} - p_0| \geq |\hat{p} - p_0|\right)$ where $\tilde{p}$ represents all possible values of $\hat{p}$. This probability is calculated using the binomial distribution.
2. $\sum_{r=0}^{X} b(r; n, p)$
3. $\sum_{r=X}^{n} b(r; n, p)$

Two approximations to these exact p-values are also given. The difference between the two is that one uses p_0 and the other uses $\hat{p}$ in the calculation of the standard error. The first approximation uses p_0 in the calculation of the standard error:

$$z_c = \frac{X + 0.5 - np_0}{\sqrt{np_0(1 - p_0)}} \text{ if } X < np_0$$

or

$$z_c = \frac{X - 0.5 - np_0}{\sqrt{np_0(1 - p_0)}} \text{ if } X > np_0$$

The second approximation uses $\hat{p}$ in the calculation of the standard error:

$$z_c = \frac{X + 0.5 - np_0}{\sqrt{n\hat{p}(1 - \hat{p})}} \text{ if } X < np_0$$

or

$$z_c = \frac{X - 0.5 - np_0}{\sqrt{n\hat{p}(1 - \hat{p})}} \text{ if } X > np_0$$

These z-values are used to calculate probabilities using the standard normal probability distribution.

Data Structure

This procedure does not use data from the database. Instead, you enter the values of n and X directly into the procedure panel.

Procedure Options

This section describes the options available in this procedure.

Data Tab

This panel specifies the data used in the analysis.

Data Values

Sample Size (n)

This is the total number of samples.

Number of Successes (X)

This is the number of successes. That is, this is the number of outcomes that exhibited the characteristic of interest.

Hypothesis Test Details

Hypothesized Proportion (P0)

This is the hypothesized proportion of successes (p_0) used in the hypothesis tests.

Alpha - Hypothesis Test

The probability in a hypothesis test of rejecting the null hypothesis (Ho) when it is true. This must be a value between 0 and 1.

Options

Alpha - Confidence Limits

The confidence coefficient to use for the confidence limits of the proportion. 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 1.

Decimal Places

The number of digits to the right of the decimal place to display.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Analysis of One Proportion

This section presents an example of how to run an analysis of data in which n is 100, X is 55, and p_0 is set to 0.50.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Proportion – One window.

1 Open the Proportion – One window.

- On the menus, select **Analysis**, then **Proportions**, then Proportion – One. The Proportion – One procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Proportion – One window, select the **Data tab**.
- In the **Sample Size (n)** box, enter **100**.
- In the **Number of Successes (X)** box, enter 55.
- In the **Hypothesized Proportion (P0)** box, enter 0.50.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Section

Data Section

Sample Size (n)	Number of Successes (X)	Sample Proportion (P)	Hypothesized Proportion (P0)	Confidence Alpha	Hypothesis Alpha
100	55.000000	0.550000	0.500000	0.050000	0.050000

This report documents the values that you gave for input.

Confidence Limits Section

Confidence Limits Section

Calculation Method	Lower 95% Confidence Limit	Sample Proportion (P)	Upper 95% Confidence Limit
Exact (Binomial)	0.447280	0.550000	0.649680
Approximation (Uncorrected)	0.452493	0.550000	0.647507
Approximation (Corrected)	0.447493	0.550000	0.652507
Wilson Score	0.452446	0.550000	0.643855

This report gives the confidence intervals. We recommend the Wilson score interval based on the article by Newcombe (1998). As a point of reference, the sample proportion, $\hat{p}$, is also given. The limits are based on the formulas that were presented earlier.

Hypothesis Test Section

Hypothesis Test Section

	Exact (Binomial)		Normal Approximation using (P0)			Normal Approximation using (P)		
Alternative Hypothesis	Prob Level	Decision (5%)	Z-Value	Prob Level	Decision (5%)	Z-Value	Prob Level	Decision (5%)
P<>P0	0.368202	Accept Ho	0.9000	0.368120	Accept Ho	0.9045	0.365712	Accept Ho
P<P0	0.864373	Accept Ho	0.9000	0.815940	Accept Ho	0.9045	0.817144	Accept Ho
P>P0	0.184101	Accept Ho	0.9000	0.184060	Accept Ho	0.9045	0.182856	Accept Ho

This report gives the results of all three hypothesis combinations. Only the alternative hypothesis is presented. The probability level and decision of each test are given. The formulas for these tests were shown earlier.

Although we present all three types of tests (two-tailed, lower-tail, and upper-tail), you would normally only select one of these tests based on your testing situation. If a choice is not obvious to you, use the two-tailed, or not equal (<>), result.

We cannot think of a reason why you would want to use the normal approximation results when then exact results are available. They were included so you could compare your answers to text books that only give approximate answers.

Chapter 515

Two Independent Proportions

Introduction

This program computes both asymptotic and exact confidence intervals and hypothesis tests for parameters used to compare two proportions. These parameters are the difference, the ratio, and the odds ratio.

Data may come from either of the following designs:

1. Random samples are drawn from two separate populations.
2. Individuals from a single sample are randomly assigned to either of two possible groups.

Comparing Two Proportions

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 a population and that the responses from one subject to the next are independent of one another.

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

	Success	Failure	Total
Population 1	*a*	*c*	*m*
Population 2	*b*	*d*	*n*
Totals	*s*	*f*	*N*

The following alternative notation is sometimes used:

	Success	Failure	Total
Population 1	x_{11}	x_{12}	n_1
Population 2	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

$$\hat{p}_1 = \frac{a}{m} = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{b}{n} = \frac{x_{21}}{n_2}$$

When analyzing studies such as these, you usually want to compare the two binomial probabilities p_1 and p_2. The most direct methods of comparing these quantities are 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

Parameter	Computation
Difference	$\delta = p_1 - p_2$
Risk Ratio	$\phi = p_1 / p_2$
Odds Ratio	$\psi = \frac{p_1 / q_1}{p_2 / q_2} = \frac{p_1 q_2}{p_2 q_1}$

The choice of which of these measures is used might at seem arbitrary, but it is important. Not only is their interpretation different, but, for small sample sizes, the decision to accept or reject the null hypothesis of no treatment effect may be different. That is, tests and confidence intervals based on these different parameters have different powers (probabilities of rejecting) and coverage probabilities. It will be useful to discuss each of these measures in detail to discover the strengths and weaknesses of each.

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 probability 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 probably 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 which 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 fits 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 for 0.1437 to 0.0793. Which single number is most 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. How do you translate from odds to probability? An odds of 3-to-2 means that the event will 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 0.60/0.4 = 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.

Another way to compare proportions is to compute the ratio of their odds. The odds ratio of two events is

$$\psi = \frac{o_1}{o_2} = \frac{\dfrac{p_1}{1 - p_1}}{\dfrac{p_2}{1 - p_2}}$$

Although the odds ratio is more complicated to interpret than the risk ratio, it is often the parameter of choice. Reasons for this include 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

$$\psi = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \approx \frac{p_1}{p_2} = \phi$$

Hypothesis Tests

A wide variety of statistical tests are available for testing hypotheses about two proportions. Some tests are based on the *difference* in proportions, others are based on the *ratio* of proportions, and still others are based on the *odds ratio*. Some tests are *conditional*, while others are *unconditional*. Some tests are said to be *large sample*, while others are said to be exact. In this section, these terms will be explained.

Types of Hypothesis Tests

Hypothesis tests concerning two proportions can be separated into three categories: large sample, conditional exact, and unconditional exact.

Large Sample Tests

Large sample (or asymptotic) tests are based on the central limit theorem (CLT) which states that for large samples, the distribution of many of these test statistics approach the normal distribution. Hence, significance levels can be computed using the normal distribution which has been extensively tabulated and can now be easily computed.

As a consequence of the asymptotic nature of the CLT, the first question that comes up is to decide when the sample size is large enough so that the CLT applies. Also, what should be done for small to medium size samples? Can these large sample tests still be used for small samples?

Exact Tests

Because of the inaccuracy of applying a large sample procedure to a small sample study, another class of tests has been devised called exact tests. The significance levels of these tests are calculated from their exact distribution, usually by considering either the binomial or the hypergeometric distribution. No appeal is made to the CLT. Because these tests involve enormous amounts of tedious calculations, they have become available with the advent of computers. In fact, specialized software programs such as StatXact have been developed to provide these calculations.

With the availability of modern computers, why aren’t approximate large sample techniques abandoned and exact tests completely embraced? To find the answer, one must delve deeper into the theory of exact tests. We will briefly summarize this here.

The distribution of the proportions in a 2-by-2 table involves two parameters: p_1 and $p_1 + \delta$ in the case of the difference and p_1 and p_1 / ϕ in the case of the ratio. The hypothesis only involves one parameter, the difference or the ratio. The other parameter, p_1, is called a *nuisance parameter* because it is not part of the hypothesis of interest. That is, the hypothesis that $\delta = 0$ or $\phi = 1$ does not involve p_1. In order to test hypotheses about the parameter of interest, the nuisance parameter must be eliminated. This may be accomplished either by conditional methods or unconditional methods.

Conditional Exact Test

The nuisance parameter can be eliminated by conditioning on a sufficient statistic. Fisher's exact test is an example of this. The conditioning occurs by considering only those tables in which the row and column totals remain the same as for the data. This removes the nuisance parameter p_1 from the distribution formula. This has drawn criticism because most experimental designs do not fix both the row and column totals. Others have argued that since the significance level is preserved unconditionally, the test is valid.

Unconditional Exact Test

The unconditional exact test approach is to remove the nuisance parameter by computing the significance level at all possible values of the nuisance parameter and choosing the largest (worst case). That is, find the value of p_1 which gives the maximum significance level (least significant) for the hypothesis test. That is, these tests find an upper bound for the significance level.

The problem with the unconditional approach is that the upper bound may occur at a value of p_1 that is far from the true value. For example, suppose the true value of p_1 is 0.711 where the significance level is 0.032. However, suppose the maximum significance level of 0.213 occurs at p_1 = 0.148. Hence, near the actual value of the nuisance value, the results are statistically significant, but the results of the exact test are not! Of course, in a particular study, we do not know the true value of the nuisance parameter. The message is that although these tests are called exact tests, they are not! They are approximate tests computed using exact distributions. Hence, you cannot say that the exact test is necessarily better than the large-sample test.

Hypotheses About the Difference in Proportions

The (risk) difference $\delta = p_1 - p_2$ is perhaps the most direct method of comparison between the 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.

The traditional approach 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 author's have begun questioning this solution. For example, based on exact enumeration, Upton (1982) and D'Agostino (1988) caution that the Fisher Exact test and Yates test should never be used.

Hypotheses About the Ratio of Proportions

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.

Hypotheses About the 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.

Large-Sample Tests

Chi-Square Test of Difference

This hypothesis test takes its place in history as one of the first statistical hypothesis tests to be proposed. It was first proposed by Karl Pearson in 1900. The two-sided test is computed as

$$\chi_1^2 = \frac{N(ad - bc)^2}{mnrs}$$

This test may also be derived as a z-test as follows

$$z = \frac{D - \delta_0}{\hat{\sigma}_D(\delta_0)}$$

where

$$D = \hat{p}_1 - \hat{p}_2$$

$$\hat{\sigma}_D(\delta_0) = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2}}$$

The quantities $\tilde{p}_1$ and $\tilde{p}_2$ are the maximum likelihood estimates constrained by $\tilde{p}_1 - \tilde{p}_2 = \delta_0$. Usually, we wish to test the hypothesis that $\delta_0 = 0$, so the standard deviation reduces to

$$\hat{\sigma}_D(0) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Chi-Square Test of Difference with Continuity Corrected

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 test was in common use when he proposed it in 1922. This test is computed as

$$\chi_1^2 = \frac{N\left(|ad - bc| - N/2\right)^2}{mnrs}$$

The continuity correction may be carried out on the z-test directly by subtracting one-half from the numerator of the test. Thus, the continuity corrected chi-square test is

$$z = \frac{|D - \delta_0| - 1/2}{\hat{\sigma}_D(\delta_0)} sign(D - \delta_0)$$

Conditional Mantel Haenszel Test and Cochran Test of Difference

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

$$E(x_{11}) = \frac{n_1 m_1}{N}$$

$$V_c(x_{11}) = \frac{n_1 n_2 m_1 m_2}{N^2(N-1)}$$

Cochran's test uses the same numerator, but a variance that has been adjusted to be unconditional. The formulas for Cochran's test are

$$z = \frac{x_{11} - E(x_{11})}{\sqrt{V_u(x_{11})}}$$

where

$$E(x_{11}) = \frac{n_1 m_1}{N}$$

$$V_u(x_{11}) = \frac{N-1}{N} V_c(x_{11})$$

Likelihood Ratio Test of Difference

In 1935, Wilks showed that the following quantity has a chi-square distribution with one degree of freedom. This test is presented, among other places, in Upton (1982). This test is computed as

$$LR = 2\left[\begin{array}{l} a\ln(a) + b\ln(b) + c\ln(c) + d\ln(d) + \\ N\ln(N) - s\ln(s) - f\ln(f) - m\ln(m) - n\ln(n) \end{array}\right]$$

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. The test is computed as

$$t_{N-2} = (ad - bc)\left(\frac{N-2}{N(nac + mbd)}\right)^{\frac{1}{2}}$$

Miettinen and Nurminen's Test of the Difference

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the difference is equal to a specified value δ_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 $N/(N\text{-}1)$ 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

$$z_{MND} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)\left(\frac{N}{N-1}\right)}}$$

where

$$\tilde{p}_1 = \tilde{p}_2 + \delta_0$$

$$\tilde{p}_2 = 2B\cos(A) - \frac{L_2}{3L_3}$$

$$A = \frac{1}{3}\left[\pi + \cos^{-1}\left(\frac{C}{B^3}\right)\right]$$

$$B = \text{sign}(C)\sqrt{\frac{L_2^2}{9L_3^2} - \frac{L_1}{3L_3}}$$

$$C = \frac{L_2^3}{27L_3^3} - \frac{L_1 L_2}{6L_3^2} + \frac{L_0}{2L_3}$$

$$L_0 = x_{21}\delta_0(1-\delta_0)$$

$$L_1 = \left[N_2\delta_0 - N - 2x_{21}\right]\delta_0 + M_1$$

$$L_2 = (N + N_2)\delta_0 - N - M_1$$

$$L_3 = N$$

Miettinen and Nurminen's Test of the Ratio

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the ratio is equal to a specified value ϕ_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.

Here is the formula for computing the test

$$z_{MNR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \phi_0^2\frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)\left(\frac{N}{N-1}\right)}}$$

where

$$\tilde{p}_1 = \tilde{p}_2\phi_0$$

$$\tilde{p}_2 = \frac{-B - \sqrt{B^2 - 4AC}}{2A}$$

$$A = N\phi_0$$

$$B = -\left[N_1\phi_0 + x_{11} + N_2 + x_{21}\phi_0\right]$$

$$C = M_1$$

Miettinen and Nurminen's 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 ψ_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

$$z_{MNO} = \frac{\frac{\left(\hat{p}_1 - \tilde{p}_1\right)}{\tilde{p}_1\tilde{q}_1} - \frac{\left(\hat{p}_2 - \tilde{p}_2\right)}{\tilde{p}_2\tilde{q}_2}}{\sqrt{\left(\frac{1}{N_2\tilde{p}_1\tilde{q}_1} + \frac{1}{N_2\tilde{p}_2\tilde{q}_2}\right)\left(\frac{N}{N-1}\right)}}$$

where

$$\tilde{p}_1 = \frac{\tilde{p}_2\psi_0}{1 + \tilde{p}_2\left(\psi_0 - 1\right)}$$

$$\tilde{p}_2 = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$$

$$A = N_2\left(\psi_0 - 1\right)$$

$$B = N_1\psi_0 + N_2 - M_1\left(\psi_0 - 1\right)$$

$$C = -M_1$$

Farrington and Manning's Test of the Difference

Farrington and Manning (1990) proposed a test statistic for testing whether the difference is equal to a specified value δ_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.

Here is the formula for computing the test

$$z_{FMD} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)}}$$

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 Test of the Ratio

Farrington and Manning (1990) proposed a test statistic for testing whether the ratio is equal to a specified value ϕ_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.

Here is the formula for computing the test

$$z_{FMR} = \frac{\hat{p}_1 / \hat{p}_2 - \phi_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \phi_0^2\frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)}}$$

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 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

$$z_{FMO} = \frac{\frac{\left(\hat{p}_1 - \tilde{p}_1\right)}{\tilde{p}_1\tilde{q}_1} - \frac{\left(\hat{p}_2 - \tilde{p}_2\right)}{\tilde{p}_2\tilde{q}_2}}{\sqrt{\left(\frac{1}{N_2\tilde{p}_1\tilde{q}_1} + \frac{1}{N_2\tilde{p}_2\tilde{q}_2}\right)}}$$

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 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 $z_{FM}(\delta)$ stand for the Farrington and Manning difference test statistic described above. The skewness corrected test statistic z_{GN} is the appropriate solution to the quadratic equation

$$\left(-\tilde{\gamma}\right)z_{GND}^2 + \left(-1\right)z_{GND} + \left(z_{FMD}(\delta) + \tilde{\gamma}\right) = 0$$

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 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 $z_{FM}(\phi)$ stand for the Farrington and Manning ratio test statistic described above. The skewness corrected test statistic z_{GN} is the appropriate solution to the quadratic equation

$$(-\tilde{\varphi})z_{GNR}^2 + (-1)z_{GNR} + (z_{FMR}(\phi) + \tilde{\varphi}) = 0$$

where

$$\tilde{\varphi} = \frac{1}{6\tilde{u}^{3/2}}\left(\frac{\tilde{q}_1(\tilde{q}_1 - \tilde{p}_1)}{n_1^2\tilde{p}_1^2} - \frac{\tilde{q}_2(\tilde{q}_2 - \tilde{p}_2)}{n_2^2\tilde{p}_2^2}\right)$$

$$\tilde{u} = \frac{\tilde{q}_1}{n_1\tilde{p}_1} + \frac{\tilde{q}_2}{n_2\tilde{p}_2}$$

Small-Sample (Exact) Tests

All of the exact tests follow the same pattern. We will present the general procedure here, and then give the specifics for each test.

General Procedure

Specify the Null and Alternative Hypotheses

The first step is to select a method to compare the proportions and determine if the test is to be one-, or two-, sided. These may be written in general as

$$H_0: h_j(p_1, p_2) = \theta_0$$

$$H_1: h_j(p_1, p_2) \neq \theta_0$$

where '$\neq$' (for two-sided tests) could be replaced with '<' or '>' for a one-sided test and the index *j* is defined as

$$h_1(p_1, p_2) = p_1 - p_2.$$

$$h_2(p_1, p_2) = p_1 / p_2$$

$$h_3(p_1, p_2) = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Specify the Reference Set

The next step is to specify the *reference set* of possible tables to compare the observed table against. Two reference sets are usually considered. Define Ω as the complete set of tables that are possible by selecting n_1 observations from one group and n_2 observations from another group. Define Γ as the subset from Ω for which $x_{11} + x_{21} = m_1$. Tests using Ω are unconditional tests while tests using Γ are conditional tests.

Specify the Test Statistic

The next step is to select the test statistic. In most cases, the score statistic is used which has the general form

$$D(y) = \frac{h_j(\hat{p}_1, \hat{p}_2) - \theta_0}{\sqrt{\tilde{V}_{h_j}(\theta_0)}}$$

where *y* represents a table with elements $y_{11}, y_{12}, y_{21}, y_{22}$ and $\tilde{V}_{h_j}(\theta_0)$ is the estimated variance of the score numerator with the constraint that the null hypothesis is true.

Select the Probability Distribution

The probability distribution an unconditional test based on the score statistic is

$$f_{p_1,p_2}(y) = \binom{n_1}{y_{11}}\binom{n_2}{y_{21}} p_1^{y_{11}} q_1^{y_{12}} p_2^{y_{21}} q_2^{y_{22}}$$

The probability distribution of a conditional test based on the score statistic is

$$f_{\psi}(y) = \frac{\binom{n_1}{y_{11}}\binom{n_2}{y_{21}} \psi^{y_{11}}}{\sum_{y \in \Gamma} \binom{n_1}{y_{11}}\binom{n_2}{y_{21}} \psi^{y_{11}}}$$

Calculate the Significance Level

The significance level (rejection probability) is found by summing the probabilities of all tables that for which the computed test statistic is at least as favorable to the alternative hypothesis as is the observed table. This may be written as

$$p(y) = \sum_{I(D(x),D(y))} f_{p_1,p_2}(y)$$

where $I(D(x), D(y))$ is an indicator function.

Maximize the Significance Level

The final step is to find the maximum value (supremum) of the significance level over all possible values of the nuisance parameter. This may be written as

$$p_{\sup\, p_2}(y) = \sup_{0<p_2<1} \left(\sum_{I(D(x),D(y))} f_{p_1,p_2}(y) \right)$$

Note that the choice of either p_1 or p_2 as the nuisance parameter is arbitrary.

Exact Tests

Barnard's Unconditional Exact Test of the Difference = 0

Barnard (1947) proposed an unconditional exact test for the difference between two proportions. It is interesting that two years later he retracted his article. However, the test has been adopted in spite of his retraction. Here are the details of this test:

Null Hypothesis: $p_1 - p_2 = 0$

Hypothesis Types: Both one-sided and two-sided

Reference Set: Ω.

Test Statistic: $$D(y) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } \hat{p} = \frac{y_{11} + y_{21}}{n_1 + n_2}$$

Two-Sided Test: $I\left(D(x), D(y)\right) = \left|D(y)\right| \geq \left|D(x)\right|$

Lower One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \leq D(x)$

Upper One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \geq D(x)$

Barnard's Exact Test of the Ratio = 1

Barnard's exact test for the ratio is identical to that for the difference.

Fisher's Conditional Exact Test of the Difference = 0

Statxact gives three conditional exact tests for testing whether the difference is zero. The most famous of these uses Fisher's statistic, but similar tests are also available using Pearson's statistic and the likelihood ratio statistic.

Null Hypothesis: $p_1 - p_2 = 0$

Hypothesis Types: Both one-sided and two-sided

Reference Set: Γ

Fisher's Test Statistic: $$D(y) = -2\ln\left(f_1(y)\right) - \ln\left(2.51N^{-3/2}\sqrt{m_1 m_2 n_1 n_2}\right)$$

L.R. Test Statistic: $$D(y) = 2\sum_{i}^{2}\sum_{j}^{2} y_{ij} \ln\left(\frac{y_{ij}}{m_i n_j / N}\right)$$

Pearson's Test Statistic: $$D(y) = \sum_{i}^{2}\sum_{j}^{2} \frac{\left(y_{ij} - m_i n_j / N\right)^2}{m_i n_j / N}$$

Two-Sided Test: $I\left(D(x), D(y)\right) = \left|D(y)\right| \geq \left|D(x)\right|$

Lower One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \le D(x)$

Upper One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \ge D(x)$

Miettinen and Nurminen's Unconditional Exact Test of the Difference

Miettinen and Nurminen (1985) proposed an unconditional exact test for testing whether the difference between two proportions is a specified value δ_0. When $\delta_0 = 0$, this test reduces to Barnard's test. Here are the details of this test:

Null Hypothesis: $p_1 - p_2 = \delta_0$

Hypothesis Types: Both one-sided and two-sided

Reference Set: Ω.

Test Statistic:
$$D(y) = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)\left(\frac{N}{N-1}\right)}}$$

where $\tilde{p}_1$ and $\tilde{p}_2$ are constrained MLE's discussed below.

Two-Sided Test: $I\left(D(x), D(y)\right) = \left|D(y)\right| \ge \left|D(x)\right|$

Lower One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \le D(x)$

Upper One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \ge D(x)$

Farrington and Manning's Unconditional Exact Test of the Difference

Farrington and Manning (1990) proposed an unconditional exact test for testing whether the difference is a specified value δ_0. This test was also discussed by Gart and Nam (1990). This test is only slightly different from the test of Miettinen and Nurminen (1985). Here are the details of this test:

Null Hypothesis: $p_1 - p_2 = \delta_0$

Hypothesis Types: Both one-sided and two-sided

Reference Set: Ω.

Test Statistic:
$$D(y) = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)}}$$

where $\tilde{p}_1$ and $\tilde{p}_2$ are constrained MLE's discussed below.

Two-Sided Test: $I\left(D(x), D(y)\right) = \left|D(y)\right| \ge \left|D(x)\right|$

Lower One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \le D(x)$

Upper One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \ge D(x)$

Miettinen and Nurminen's Unconditional Exact Test of the Ratio

Miettinen and Nurminen (1985) proposed an unconditional exact test for testing whether the ratio between two proportions is a specified value ϕ_0. When $\phi_0 = 1$, this test reduces to Barnard's test. Here are the details of this test:

Null Hypothesis: $p_1 / p_2 = \phi_0$

Hypothesis Types: Both one-sided and two-sided

Reference Set: Ω.

Test Statistic: $$D(y) = \frac{\hat{p}_1 - \hat{p}_2 - \phi_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \phi_0^2\frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)\left(\frac{N}{N-1}\right)}}$$

Two-Sided Test: $I\left(D(x), D(y)\right) = \left|D(y)\right| \ge \left|D(x)\right|$

Lower One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \le D(x)$

Upper One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \ge D(x)$

Farrington and Manning's Unconditional Exact Test of the Ratio

Farrington and Manning (1990) proposed an unconditional exact test for testing whether the ratio is a specified value ϕ_0. This test was also discussed by Gart and Nam (1988). This test is only slightly different from the test of Miettinen and Nurminen (1985). Here are the details of this test:

Null Hypothesis: $p_1 / p_2 = \phi_0$

Hypothesis Types: Both one-sided and two-sided

Reference Set: Ω.

Test Statistic: $$D(y) = \frac{\hat{p}_1 - \hat{p}_2 - \phi_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \phi_0^2\frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)}}$$

Two-Sided Test: $I\left(D(x), D(y)\right) = \left|D(y)\right| \ge \left|D(x)\right|$

Lower One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \le D(x)$

Upper One-Sided Test: $I\left(D(x), D(y)\right) = D(y) \ge D(x)$

Constrained MLE's

The Miettinen and Nurminen (1985) and Farrington and Manning (1990) tests given above require maximum likelihood estimates that are constrained to follow the null hypothesis that $p_1 - p_2 = \delta_0$. The constrained maximum likelihood estimate for $\tilde{p}_2$ when considering the difference is the appropriate solution of the cubic equation

$$N\tilde{p}_2^3 + \left[(N + n_2)\delta_0 - (N + m_2)\right]\tilde{p}_2^2 + \left[m_2 - \delta_0(N + 2y_{21}) + n_2\delta_0^2\right]\tilde{p}_2 + y_{21}\delta_0(1 - \delta_0) = 0$$

The value for $\tilde{p}_1$ is found using the constraint

$$\tilde{p}_1 = \tilde{p}_2 + \delta_0$$

The constrained maximum likelihood estimate of $\tilde{p}_2$ when considering the ratio with the constraint that $p_1 / p_2 = \phi_0$ is the appropriate solution of the quadratic equation

$$N\tilde{p}_2^2 - \left[\phi_0(n_1 + y_{11}) + y_{21} + n_1\right]\tilde{p}_2 + m_2 = 0$$

The value for $\tilde{p}_1$ is found using the constraint

$$\tilde{p}_1 = \tilde{p}_2\phi_0$$

Equivalence Tests for the Difference and Ratio

An equivalence test is designed to show that one (new) treatment is similar to, but not necessarily better than, another (standard) treatment. To accomplish this, the roles of the null and alternative hypotheses are reversed. The hypotheses for testing equivalence of the difference in proportions are (assuming that $\delta_L < 0$ and $\delta_U > 0$)

$$H_0: p_1 - p_2 \le \delta_L \quad \text{or} \quad p_1 - p_2 \ge \delta_U \quad \text{versus} \quad H_1: \delta_L < p_1 - p_2 < \delta_u$$

The hypotheses for testing equivalence of the ratio of proportions are (assuming that $\phi_L < 1$ and $\phi_U > 1$)

$$H_0: p_1 / p_2 \le \phi_L \quad \text{or} \quad p_1 / p_2 \ge \phi_U \quad \text{versus} \quad H_1: \phi_L < p_1 / p_2 < \phi_u$$

The alternative hypothesis states that the true value is between some small, clinically acceptable range. For example, we might be willing to conclude that the benefits of two drugs are equivalent if the difference in their response rates is between -0.05 and 0.05 or if the ratio of their response rates is between 0.90 and 1.10.

The conventional method of testing equivalence hypotheses is to perform two, one-sided tests (TOST) of hypotheses. The null hypothesis of non-equivalence is rejected in favor of the alternative hypothesis of equivalence if both one-sided tests are rejected. Unlike the common two-sided tests, however, the type I error rate is set directly at the nominal level (usually 0.05)—it is not split in half. So, to perform the test, two, one-sided tests are conducted at the α significance level. If both are rejected, the alternative hypothesis is concluded at the α significance level. Note that the *p*-value of the test is the maximum of the *p*-values of the two tests.

The two, one-sided tests of hypotheses for the difference are

$$H_{01}: p_1 - p_2 \le \delta_L \quad \text{versus} \quad H_{11}: p_1 - p_2 > \delta_L$$

$$H_{02}: p_1 - p_2 \ge \delta_U \quad \text{versus} \quad H_{12}: p_1 - p_2 < \delta_U$$

The two, one-sided tests of hypotheses for the ratio are

$$H_{01}: p_1 / p_2 \le \phi_L \quad \text{versus} \quad H_{11}: p_1 / p_2 > \phi_L$$

$$H_{02}: p_1 / p_2 \ge \phi_U \quad \text{versus} \quad H_{12}: p_1 / p_2 < \phi_U$$

These one-sided tests can use any of the large-sample or exact one-sided tests that were discussed earlier.

Confidence Intervals

Both large sample and exact confidence intervals may be computed for the difference, the ratio, and the odds ratio.

Confidence Intervals for the Difference

Several methods are available for computing a confidence interval of the difference between two proportions $\delta = p_1 - p_2$. Newcombe (1998) conducted a comparative evaluation of eleven confidence interval methods. He recommended that the modified Wilson score method be used instead of the Pearson Chi-Square or the Yate's Corrected Chi-Square. Beal (1987) found that the Score methods performed very well. The lower *L* and upper *U* limits of these intervals are computed as follows. Note that, unless otherwise stated, $z = |z_{\alpha/2}|$ is the appropriate percentile from the standard normal distribution.

Cells with Zero Counts

Extreme cases in which some cells are zero require special approaches with some of the tests given below. We have found that a simple solution that works well is to change the zeros to a small positive number such as 0.01. This produces the same results as other techniques that we are aware.

C.I. for Difference: Pearson's Chi-Square

For details, see Newcombe (1998), page 875.

$$L = \hat{p}_1 - \hat{p}_2 - z\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}\right)}$$

$$U = \hat{p}_1 - \hat{p}_2 + z\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}\right)}$$

C.I. for Difference: Yate's Chi-Square with Continuity Correction

For details, see Newcombe (1998), page 875.

$$L = \hat{p}_1 - \hat{p}_2 - z\left[\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}\right)} + \frac{1}{2}\left(\frac{1}{m} + \frac{1}{n}\right)\right]$$

$$U = \hat{p}_1 - \hat{p}_2 + z\left[\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}\right)} + \frac{1}{2}\left(\frac{1}{m} + \frac{1}{n}\right)\right]$$

C.I. for Difference: Wilson's Score as modified by Newcombe

For details, see Newcombe (1998), page 876.

$$L = \hat{p}_1 - \hat{p}_2 - B$$

$$U = \hat{p}_1 - \hat{p}_2 + C$$

where

$$B = z\sqrt{\frac{l_1(1-l_1)}{m} + \frac{u_2(1-u_2)}{n}}$$

$$C = z\sqrt{\frac{u_1(1-u_1)}{m} + \frac{l_2(1-l_2)}{n}}$$

and l_1 and u_1 are the roots of

$$\left|p_1 - \hat{p}_1\right| - z\sqrt{\frac{p_1(1-p_1)}{m}} = 0$$

and l_2 and u_2 are the roots of

$$\left|p_2 - \hat{p}_2\right| - z\sqrt{\frac{p_2(1-p_2)}{n}} = 0$$

C.I. for Difference: Farrington and Manning's Score

Farrington and Manning (1990) proposed inverting their score test to find the confidence interval. The details of calculating their score statistic *z* are presented above in the Hypothesis Test section. The lower limit is found by solving

$$z_{FMD} = \left|z_{\alpha/2}\right|$$

and the upper limit is the solution of

$$z_{FMD} = -\left|z_{\alpha/2}\right|$$

C.I. for Difference: Miettinen and Nurminen's Score

Miettinen and Nurminen (1985) proposed inverting their score test to find the confidence interval. The details of calculating their score statistic *z* are presented above in the Hypothesis Test section. The lower limit is found by solving

$$z_{MND} = \left|z_{\alpha/2}\right|$$

and the upper limit is the solution of

$$z_{MND} = -\left|z_{\alpha/2}\right|$$

C.I. for Difference: Gart and Nam's Score

Gart and Nam (1988) proposed inverting their score test to find the confidence interval. The details of calculating their score statistic *z* are presented above in the Hypothesis Test section. The lower limit is found by solving

$$z_{GND} = \left|z_{\alpha/2}\right|$$

and the upper limit is the solution of

$$z_{GND} = -\left|z_{\alpha/2}\right|$$

C.I. for Difference: Chen's Quasi-Exact Method

Chen (2002) proposed a quasi-exact method for generating confidence intervals. This method produces intervals that are close to unconditional exact intervals that are available in specialized software like StatXact, but do not require as much time to compute. Chen's method inverts a hypothesis test based on Farrington and Manning's method. That is, the confidence interval is found by finding those values at which the hypothesis test that the difference is a given, non-zero value become significant. However, instead of searching for the maximum significance level of all possible values of the nuisance parameter as the exact tests do, Chen proposed using the significance level at the constrained maximum likelihood estimate of p_2 as given by Farrington and Manning. This simplification results in a huge reduction in computation with only a minor reduction in accuracy. Also, it allows much larger sample sizes to be analyzed.

Note on Exact Tests

A word of caution should be raised about the phrase *exact tests*. Many users assume that methods that are based on exact methods are always better than other, non-exact methods. After all, 'exact' sounds better than 'approximate'. However, tests and confidence intervals based on exact methods are not necessarily better. In fact, some prominent statisticians are of the opinion that they are actually worse! (See Agresti and Coull (1998) for one example). *Exact* simply means that they are based on exact distributional calculations. They are, however, conservative in terms of their coverage probabilities (the probability that the confidence interval includes the true value). That is, they are wider than they need to be because they are based on worst case scenarios. So the bottom line is this—do not always assume that exact methods are the better methods.

Confidence Intervals for the Ratio

C.I. for Ratio: Farrington and Manning's Score

Farrington and Manning (1990) proposed inverting their score test to find the confidence interval. The details of calculating their score statistic *z* are presented above in the Hypothesis Test section. The lower limit is found by solving

$$z_{FMR} = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z_{FMR} = -|z_{\alpha/2}|$$

C.I. for Ratio: Miettinen and Nurminen's Score

Miettinen and Nurminen (1985) proposed inverting their score test to find the confidence interval. The details of calculating their score statistic *z* are presented above in the Hypothesis Test section. The lower limit is found by solving

$$z_{MNR} = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z_{MNR} = -|z_{\alpha/2}|$$

C.I. for Ratio: Gart and Nam's Score

Gart and Nam (1988) proposed inverting their score test to find the confidence interval. The details of calculating their score statistic *z* are presented above in the Hypothesis Test section. The lower limit is found by solving

$$z_{GNR} = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z_{GNR} = -|z_{\alpha/2}|$$

C.I. for Ratio: Logarithm (Katz)

This was one of the first methods proposed for computing confidence intervals for risk ratios.

For details, see Gart and Nam (1988), page 324.

$$L = \hat{\phi}\exp\left(-z\sqrt{\frac{\hat{q}_1}{n\hat{p}_1}+\frac{\hat{q}_2}{n\hat{p}_2}}\right)$$

$$U = \hat{\phi}\exp\left(z\sqrt{\frac{\hat{q}_1}{n\hat{p}_1}+\frac{\hat{q}_2}{n\hat{p}_2}}\right)$$

where

$$\hat{\phi} = \frac{\hat{p}_1}{\hat{p}_2}$$

C.I. for Ratio: Logarithm (Walters)

For details, see Gart and Nam (1988), page 324.

$$L = \hat{\phi}\exp\left(-z\sqrt{\hat{u}}\right)$$

$$U = \hat{\phi}\exp\left(z\sqrt{\hat{u}}\right)$$

where

$$\hat{\phi} = \exp\left(\ln\left(\frac{a+\frac{1}{2}}{m+\frac{1}{2}}\right) - \ln\left(\frac{b+\frac{1}{2}}{n+\frac{1}{2}}\right)\right)$$

$$\hat{u} = \frac{1}{a+\frac{1}{2}} - \frac{1}{m+\frac{1}{2}} + \frac{1}{b+\frac{1}{2}} - \frac{1}{n+\frac{1}{2}}$$

$$\tilde{q}_2 = 1 - \tilde{p}_2$$

$$V = \left(\phi^2\left(\frac{\tilde{q}_1}{m\tilde{p}_1} + \frac{\tilde{q}_2}{n\tilde{p}_2}\right)\right)^{-1}$$

$$\tilde{p}_1 = \phi\tilde{p}_2$$

$$\tilde{q}_1 = 1 - \tilde{p}_1$$

$$\tilde{q}_2 = 1 - \tilde{p}_2$$

$$\tilde{\mu}_3 = v^{3/2}\left(\frac{\tilde{q}_1(\tilde{q}_1 - \tilde{p}_1)}{(m\tilde{p}_1)^2} - \frac{\tilde{q}_2(\tilde{q}_2 - \tilde{p}_2)}{(n\tilde{p}_2)^2}\right)$$

$$v = \left(\frac{\tilde{q}_1}{m\tilde{p}_1} + \frac{\tilde{q}_2}{n\tilde{p}_2}\right)^{-1}$$

C.I. for Ratio: Chen's Quasi-Exact Method

Chen (2002) proposed a quasi-exact method for generating confidence intervals. This method produces intervals that are close to unconditional exact intervals that are available in specialized software like StatXact, but do not require as much time to compute. Chen's method inverts a hypothesis test based on Farrington and Manning's method. That is, the confidence interval is found by finding those values at which the hypothesis test that the difference is a given, non-zero value become significant. However, instead of searching for the maximum significance level of all possible values of the nuisance parameter as the exact tests do, Chen proposed using the significance level at the constrained maximum likelihood estimate of p_2 as given by Farrington and Manning. This simplification results in a huge reduction in computation with only a minor reduction in accuracy. Also, it allows much larger sample sizes to be analyzed.

Confidence Intervals for the Odds Ratio

The odds ratio is a commonly used measure of treatment effect when comparing two binomial proportions. It is the ratio of the odds of the event in group one divided by the odds of the event in group two. The results given below are found in Fleiss (1981).

Symbolically, the odds ratio is defined as

$$\psi = \frac{\dfrac{p_1}{1-p_1}}{\dfrac{p_2}{1-p_2}}$$

C.I. for Odds Ratio: Simple Technique

The simple estimate of the odds ratio uses the formula

$$\hat{\psi} = \frac{\hat{p}_1\hat{q}_2}{\hat{p}_2\hat{q}_1} = \frac{ad}{bc}$$

The standard error of this estimator is estimated by

$$se(\hat{\psi}) = \hat{\psi}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

Problems occur if any one of the quantities *a*, *b*, *c*, or *d* are zero. To correct this problem, many authors recommend adding one-half to each cell count so that a zero cannot occur. Now, the formulas become

$$\hat{\psi}' = \frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)}$$

and

$$se(\hat{\psi}') = \hat{\psi}'\sqrt{\frac{1}{a+0.5}+\frac{1}{b+0.5}+\frac{1}{c+0.5}+\frac{1}{d+0.5}}$$

The distribution of these direct estimates of the odds ratio do not converge to normality as fast as does their logarithm, so the logarithm of the odds ratio is used to form confidence intervals. The formula for the standard error of the log odds ratio is

$$L' = \ln\left(\hat{\psi}'\right)$$

and

$$se\left(L'\right) = \sqrt{\frac{1}{a+0.5} + \frac{1}{b+0.5} + \frac{1}{c+0.5} + \frac{1}{d+0.5}}$$

A $100(1-\alpha)\%$ confidence interval for the log odds ratio is formed using the standard normal distribution as follows

$$\hat{\psi}_{lower} = \exp\left(L' - z_{1-\alpha/2}se\left(L'\right)\right)$$

$$\hat{\psi}_{upper} = \exp\left(L' + z_{1-\alpha/2}se\left(L'\right)\right)$$

C.I. for Odds Ratio: Iterated Method of Fleiss

Fleiss (1981) presents an improve confidence interval for the odds ratio. This method forms the confidence interval as all those value of the odds ratio which would not be rejected by a chi-square hypothesis test. Fleiss gives the following details about how to construct this confidence interval. To compute the lower limit, do the following.

1. For a trial value of ψ, compute the quantities *X*, *Y*, *W*, *F*, *U*, and *V* using the formulas

$$X = \psi\left(m+s\right) + \left(n-s\right)$$

$$Y = \sqrt{X^2 - 4ms\psi\left(\psi-1\right)}$$

$$A = \frac{X-Y}{2\left(\psi-1\right)}$$

$$B = s - A$$

$$C = m - A$$

$$D = f - m + A$$

$$W = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

$$F = \left(a - A - \tfrac{1}{2}\right)^2 W - z_{\alpha/2}^2$$

$$T = \frac{1}{2\left(\psi-1\right)^2}\left(Y - n_{\cdot\cdot} - \frac{\psi-1}{Y}\left[X\left(m+s\right) - 2ms\left(2\psi-1\right)\right]\right)$$

$$U = \frac{1}{B^2} + \frac{1}{C^2} - \frac{1}{A^2} - \frac{1}{D^2}$$

$$V = T\left[\left(a - A - \tfrac{1}{2}\right)^2 U - 2W\left(a - A - \tfrac{1}{2}\right)\right]$$

Finally, use the updating equation below to calculate a new value for the odds ratio using the updating equation

$$\psi^{(k+1)} = \psi^{(k)} - \frac{F}{V}$$

2. Continue iterating until the value of F is arbitrarily close to zero.

The upper limit is found by substituting $+\frac{1}{2}$ for $-\frac{1}{2}$ in the formulas for F and V.

Confidence limits for the *relative risk* can be calculated using the expected counts A, B, C, and D from the last iteration of the above procedure. The lower limit of the relative risk

$$\phi_{lower} = \frac{A_{lower} n}{B_{lower} m}$$

$$\phi_{upper} = \frac{A_{upper} n}{B_{upper} m}$$

C.I. for Odds Ratio: Mantel-Haenszel

The common estimate of the logarithm of the odds ratio is used to create this estimator. That is

$$\ln\left(\hat{\psi}\right) = \ln\left(\frac{ad}{bc}\right)$$

The standard error of this estimator is estimated using the Robins, Breslow, Greenland (1986) estimator which performs well in most situations. The standard error is given by

$$\text{se}\left(\ln\left(\hat{\psi}\right)\right) = \sqrt{\frac{A}{2C} + \frac{AD + BC}{2CD} + \frac{B}{2D}}$$

where

$$A = \frac{a + d}{N}$$

$$B = \frac{b + c}{N}$$

$$C = \frac{ad}{N}$$

$$D = \frac{bc}{N}$$

The confidence limits are calculated as

$$\hat{\psi}_{lower} = \exp\left(\ln\left(\hat{\psi}\right) - z_{1-\alpha/2} se\left(\ln\left(\hat{\psi}\right)\right)\right)$$

$$\hat{\psi}_{upper} = \exp\left(\ln\left(\hat{\psi}\right) + z_{1-\alpha/2} se\left(\ln\left(\hat{\psi}\right)\right)\right)$$

C.I. for Odds Ratio: Conditional Exact

The conditional exact confidence interval of the odds ratio is calculated using the noncentral hypergeometric distribution as given in Sahai and Khurshid (1995). That is, a $100(1-\alpha)\%$ confidence interval is found by searching for ψ_L and ψ_U such that

$$\frac{\sum_{k=x}^{k_2}\binom{n_1}{k}\binom{n_2}{m_1-k}(\psi_L)^k}{\sum_{k=k_1}^{k_2}\binom{n_1}{k}\binom{n_2}{m_1-k}(\psi_L)^k}=\frac{\alpha}{2}$$

$$\frac{\sum_{k=k_1}^{x}\binom{n_1}{k}\binom{n_2}{m_1-k}(\psi_U)^k}{\sum_{k=k_1}^{k_2}\binom{n_1}{k}\binom{n_2}{m_1-k}(\psi_U)^k}=\frac{\alpha}{2}$$

where

$$k_1=\max(0,m_1-n_1) \text{ and } k_2=\min(n_1,m_1)$$

Data Structure

This procedure does not use data from the database. Instead, you enter the values directly into the panel. The data are entered in the familiar 2-by-2 table format.

Procedure Options

This section describes the options available in this procedure.

Data Tab

Enter the data values directly on this panel.

Data Values

A Count (Group = 1, Response = Positive)

This is the count for cell *A* of the 2-by-2 table. This is the number in group 1 that had a positive response.

B Count (Group = 2, Response = Positive)

This is the count for cell *B* of the 2-by-2 table. This is the number in group 2 that had a positive response.

C Count (Group = 1, Response = Negative)

This is the count for cell *C* of the 2-by-2 table. This is the number in group 1 that had a negative response.

D Count (Group = 2, Response = Negative)

This is the count for cell *D* of the 2-by-2 table. This is the number in group 2 that had a negative response.

Statistic(s)

These three boxes indicate which of the possible test statistics (difference, ratio, or odds ratio) are to be displayed in the reports. At least one statistic must be checked.

Confidence Intervals

These three boxes indicate which of the possible types of confidence intervals (large sample, exact, or bootstrap) are to be displayed in the reports. Note that exact confidence intervals often take a long time to calculated, especially with large ($N > 100$) counts.

Hypothesis Tests – Tests

These two boxes indicate whether you want the large-sample and/or the exact tests displayed. If neither box is checked, no hypothesis tests will be computed.

Hypothesis Tests – Null Hypothesis

These two boxes indicate the type of hypothesis tests that are to be run. If neither box is checked, no hypothesis tests will be computed.

H0 = 0 for Diff's and 1 for Ratios

Check this option to run the standard tests in which the null hypothesis is that the proportion difference is zero, the ratio is one, or the odds ratio is one.

H0 = User-Specified Value

Check this option to run the hypothesis tests in which the null hypothesis is that the test statistic is equal to the corresponding value given under H0 Values tab. Note that this option is checked when you want to run a noninferiority test.

Hypothesis Tests – Alternative Hypothesis

Check these boxes to indicate which of the possible hypothesis tests are to be run.

Confidence Intervals Tab

This panel contains options that control the confidence intervals.

Confidence Interval Options

Confidence Interval Alpha

This option sets the alpha value for any confidence limits that are generated. The confidence coefficient of a confidence interval is equal to 1 - alpha. Thus, an alpha of 0.05 results in a confidence coefficient of 95%. Typical values are 0.01, 0.05, and 0.10.

Confidence Interval Options - Bootstrap Confidence Interval Options

Bootstrap Samples

This is the number of bootstrap samples used. A general rule of thumb is that you use at least 100 when standard errors are your focus or at least 1000 when confidence intervals are your focus. If computing time is available, it does not hurt to do 4000 or 5000.

We recommend setting this value to at least 3000.

Percentile Type

The method used to create the percentiles when forming bootstrap confidence limits. You can read more about the various types of percentiles in the Descriptive Statistics chapter. We suggest you use the Ave *X*(p[n+1]) option.

C.I. Method

This option specifies the method used to calculate the bootstrap confidence intervals. The reflection method is recommended.

- **Percentile**

  The confidence limits are the corresponding percentiles of the bootstrap values.

- **Reflection**

  The confidence limits are formed by reflecting the percentile limits. If *X*0 is the original value of the parameter estimate and *XL* and *XU* are the percentile confidence limits, the Reflection interval is (2 *X*0 - *XU*, 2 *X*0 - *XL*).

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. We recommend setting the parameter to at least 50.

Random Number Seed

Use this option to specify the seed value of the random number generator. Specify a number between 1 and 32000 to seed (start) the random number generator. This seed will be used to start the random number generator, so you will obtain the same results whenever it is used.

If you want to have a random start, enter the phrase 'RANDOM SEED'.

Hypothesis Tests Tab

This panel contains options that control the hypothesis tests.

H0 User-Specified Value when 'H0 = User-Specified Value is checked

Difference

Enter the hypothesized value of the difference between proportions δ_0 under the null hypothesis. The standard hypothesis tests test whether this value is zero. This option lets you specify a value other than zero, which is commonly used for noninferiority tests (see the noninferiority example for more details).

The possible range of values for the difference is between -1 and 1.

This option is only used when the 'H0 = User Specified Value' option is checked.

Ratio

Enter the hypothesized value of the ratio of the proportions ϕ_0 under the null hypothesis. The standard hypothesis tests test whether the ratio is one. This option lets you specify a value other than one.

This option is only used when the 'H0 = User Specified Value' option is checked.

The possible range of values for the ratio is any positive number. Usually, a number between 0.25 and 4.0 is used.

Odds Ratio

Enter the hypothesized value of the odds ratio ψ_0 under the null hypothesis. The standard hypothesis tests test whether the odds ratio is one. This option lets you specify a value other than one.

This option is only used when the 'H0 = User Specified Value' option is checked.

The possible range of values for the ratio is any positive number. Usually, a number between 0.25 and 4.0 is used.

Exact Tests

The options on this panel control which test statistics are used in the exact hypothesis tests and confidence intervals.

D R O

These check boxes control which test statistics are used. The three choices below 'D' are for differences, the three choices below 'R' are for ratios, and the two choices below 'O' are for odds ratios. Usually, only one box in each column needs to be checked.

The possible choices are uncorrected tests of Farrington and Manning, the corrected tests of Miettenen and Numinen, and the skewness corrected tests of Gart and Nam. Note that the defaults match the tests available in StatXact.

There is little reason to run all tests since the results are usually identical.

Equivalence Bounds

Lower and Upper Equivalence Bounds for Difference

These options specify the upper and lower equivalence bounds for the equivalence tests of the difference in proportions. That is, these options specify δ_U and δ_L. Usually, $\delta_L = -\delta_U$, but this is not required.

This value is sometimes called the *margin of equivalence*. They represent the largest difference that would still result in the conclusion of equivalence. For example, suppose that if response rates of two drugs are 0.71 and 0.79, they are considered equivalent. However, if the two rates are 0.71 and 0.80, they are not considered equivalent. Then, in this case, the margin of equivalence is 0.09.

The possible range of values is between 0 and 1. Typical values are between 0.05 and 0.25.

This option is only used when the 'Equivalence' hypothesis tests option is checked.

Lower and Upper Equivalence Bounds for Ratio

These options specify the upper and lower equivalence bounds for the equivalence tests of the difference in proportions. That is, these options specify ϕ_U and ϕ_L. Usually, $\phi_L = 1/\phi_U$, but this is not required.

This value is sometimes called the *margin of equivalence*. They represent the largest difference that would still result in the conclusion of equivalence. The possible range of values is between zero and one for ϕ_L and greater than one for ϕ_U. Typical values are between 0.50 and 2.0.

This option is only used when the 'Equivalence' hypothesis tests option is checked.

Lower and Upper Equivalence Bounds for Odds Ratio

These options specify the upper and lower equivalence bounds for the equivalence tests of the difference in proportions. That is, these options specify ψ_U and ψ_L. Usually, $\psi_L = 1/\psi_U$, but this is not required.

This value is sometimes called the *margin of equivalence*. They represent the largest difference that would still result in the conclusion of equivalence. The possible range of values is between zero and one for ψ_L and greater than one for ψ_U. Typical values are between 0.50 and 2.0.

This option is only used when the 'Equivalence' hypothesis tests option is checked.

Alpha

Hypothesis Test Alpha

The probability in a hypothesis test of rejecting the null hypothesis (H0) when it is true.

Reports Tab

This panel contains options that control the contents and format of the output.

Data Summary Reports

Original Data Summary Report

Check this option to display a report of the original data values and resulting proportions.

Report Decimal Places

Decimal Places – Proportions to Test Values

The number of digits to the right of the decimal place to display for each type of value.

Options Tab

This panel contains miscellaneous options about the computations.

Limits for Exact Results

Maximum N for Exact

Specify the maximum allowable value of $N = N1+N2$ for exact confidence intervals and hypothesis tests. When N is greater than this amount, the 'exact' procedures are not calculated.

Because of the long running time needed for $N>50$, this option lets you set a limit. Since the score test results are very close to the exact test results for larger ($N>100$) sample sizes, there is little point in spending the time to calculate exact procedures for ‘large’ samples.

Maximum N for Quasi-Exact

Specify the maximum allowable value of $N=N1+N2$ for quasi-exact confidence intervals. When N is greater than this amount, the quasi-exact confidence intervals are not calculated.

Because of the long running time needed for $N>100$, this option lets you set a limit. Since the intervals based on the score test are very close to the exact test results for larger ($N>100$) sample sizes, there is little point in spending the time to calculate quasi-exact confidence intervals for ‘large’ samples.

Number of Intervals in Searches

Specify the number of intervals to be used in the grid searches used in the exact tests and exact confidence intervals. Usually, ‘40’ will obtain answers correct to three places. For tables with an $N > 100$, you may want to reduce this to 20 because of the lengthy compute time.

Zero Count Adjustment

Zero Count Adjustment Method

Zero cell counts cause many calculation problems with ratios and odds ratios. To compensate for this, a small value (called the Zero Adjustment Value) may 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.

Adding a small value is controversial, but may be necessary. Some statisticians recommend adding 0.5 while others recommend 0.25. We have found that adding values as small as 0.001 seems to work well. However, you may have to defend your choice, so when possible, do not add anything.

Zero Count Adjustment Value

Zero cell counts cause many calculation problems. 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.

Some statisticians recommend that the value of 0.5 be added to all cells (both zero and non-zero). Others recommend 0.25. We have found that even a value as small as 0.01 works well.

The value of the ratio and the odds ratio will depend on the amount chosen here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Large-Sample Analysis

This section presents an example of a standard, large-sample analysis of the difference between two proportions. In this example, 14 of 26 receiving the standard treatment responded positively and 19 of 23 receiving the experimental treatment responded positively.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Proportions – Two Independent window.

1 Open the Proportions – Two Independent procedure.

- On the menus, select **Analysis**, then **Proportions**, then **Proportions – Two Independent**. The Proportions – Two Independent procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Proportions – Two Independent window, select the **Data tab**.
- In the **A Count** box, enter **19**.
- In the **C Count** box, enter **4**.
- In the **B Count** box, enter **14**.
- In the **D Count** box, enter **12**.
- Under the Statistics heading, check **Difference (P1 – P2)**, **Ratio (P1/P2)**, and **Odds Ratio [P1/(1-P1)] / [P2/(1-P2)]**. Note that usually, you would only need to check one of these.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, click the Run button (the left-most button on the button bar at the top).

Table and Data Sections

Table Section

A	B	C	D	N1 (A+C)	N2 (B+D)	M1 A+B	M2 (C+D)	N (N1+N2)
19	14	4	12	23	26	33	16	49

Data Section

Sample	Sample Size	Positive Responses	Negative Responses	Proportion Positive	Proportion Negative
One	23	19	4	0.8261	0.1739
Two	26	14	12	0.5385	0.4615
Total	49	33	16	0.6735	0.3265

This report documents the values that were input.

Confidence Intervals

Confidence Intervals of Difference (P1-P2)

Confidence Interval Method	Estimated Value	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Score (Farrington & Manning)	0.2876	0.0255	0.5124
Score (Miettinen & Nurminen)	0.2876	0.0227	0.5145
Score w/Skewness (Gart-Nam)	0.2876	0.0268	0.5196
Score (Wilson)	0.2876	0.0245	0.4989
Score (Wilson C.C.)	0.2876	-0.0044	0.5200
Chi-Square C.C. (Yates)	0.2876	0.0003	0.5750
Chi-Square (Pearson)	0.2876	0.0412	0.5340

Confidence Intervals of Ratio (P1/P2)

Confidence Interval Method	Estimated Value	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Score (Farrington-Manning)	1.53	1.04	2.40
Score (Miettinen-Nurminen)	1.53	1.03	2.41
Score w/Skewness (Gart-Nam)	1.53	1.04	2.43
Logarithm (Katz)	1.53	1.03	2.29
Logarithm + 1/2 (Walter)	1.52	1.02	2.24
Fleiss	1.53	0.98	2.13

Confidence Intervals of Odds Ratio (Odds1/Odds2)

Confidence Interval Method	Estimated Value	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Exact (Conditional)	3.74	0.94	20.52
Score (Farrington & Manning)	4.07	1.12	14.56
Score (Miettinen & Nurminen)	4.07	1.11	14.74
Fleiss's Iterated	3.74	0.93	19.16
Logarithmic	3.74	1.04	13.35
Mantel-Haenszel	4.07	1.08	15.33
Simple	4.07	0.00	9.47
Simple + 1/2	3.74	0.00	8.49

These reports provide large sample confidence intervals based on formulas shown earlier in this chapter. The first interval in each list is recommended. Unless you have a good reason for selecting another interval, you should use this one.

The interpretation of these confidence intervals is that when populations are repeatedly sampled and confidence intervals are calculated, 95% of those confidence intervals will include (cover) the true value of the parameter (actual difference, ratio, or odds ratio).

Hypothesis Tests

Two-Sided Tests of Zero Difference (H0: P1 = P2 versus H1: P1 <> P2)
Estimated Difference (P1 - P2) = 0.2876

Test Name	Test Statistic's Distribution	Test Statistic Value	Prob Level	Conclude H1 at 5% Significance?
Fisher's Exact	Hypergeometric		0.0388	Yes
Chi-Square Test	Chi-Square(1)	4.591	0.0321	Yes
Chi-Square Test (C.C.)	Chi-Square(1)	3.376	0.0661	No
Z-Test	Normal	2.143	0.0321	Yes
Z-Test (C.C.)	Normal	1.837	0.0661	No
Mantel-Haenszel Test	Normal	2.121	0.0339	Yes
Likelihood Ratio	Chi-Square(1)	4.763	0.0291	Yes
T-Test using 0's and 1's	Student's T(47)	1.963	0.0556	No

Two-Sided Tests of Ratio Unity (H0: P1/P2 = 1 versus H1: P1/P2 <> 1)
Estimated Ratio (P1 / P2) = 1.534

Test Name	Test Statistic's Distribution	Test Statistic Value	Prob Level	Conclude H1 at 5% Significance?
Z-Test	Normal	2.143	0.0321	Yes

Two-Sided Tests of Odds Ratio Unity (H0: Odds1 / Odds2 = 1 versus H1: Odds1 / Odds2 <> 1)
Estimated Odds Ratio (Odds1 / Odds2) = 4.071

Test Name	Test Statistic's Distribution	Test Statistic Value	Prob Level	Conclude H1 at 5% Significance?
Exact	Hypergeometric	19.000	0.0637	No
Ln(Odds Ratio)	Normal	2.076	0.0379	Yes
Mantel-Haenszel	Normal	2.076	0.0379	Yes

These reports give the result of several large-sample tests of the hypothesis. The formulas for these tests were shown earlier.

Although many tests are provided for the difference and for the odds ratio, you should use only one. In fact, you should have picked the test statistic you are going to use before running the study. It is inappropriate to scan through the results and select a test that matches your desired conclusion.

Test Name

This column gives the name of the test.

Test Statistic’s Distribution

This column gives the name of the distribution that the test statistic follows under the null hypothesis.

Test Statistic Value

This is the value of the test statistic under the null hypothesis.

Prob Level

This is the significance level of the test. When this value is less than the critical value (often 0.05), the test is ‘significant’. Otherwise, it is not.

Conclude H1 at 5% Significance?

The column indicates whether the test is significant or not at the indicated significance level.

Example 2 – Exact Tests and Intervals

This section presents an example of an exact test of the difference between the two proportions. In this example, 14 of 26 receiving the standard treatment responded positively and 19 of 23 receiving the experimental treatment responded positively.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Proportions – Two Independent window.

1 Open the Proportions – Two Independent procedure.

- On the menus, select **Analysis**, then **Proportions**, then **Proportions – Two Independent**. The Proportions – Two Independent procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Proportions – Two Independent window, select the **Data tab**.
- In the **A Count** box, enter **19**.
- In the **C Count** box, enter **4**.
- In the **B Count** box, enter **14**.
- In the **D Count** box, enter **12**.
- Under Statistic(s), check only the **Difference (P1-P2)** box.
- Under Confidence Intervals, check **Exact** (in addition to Large-Sample).
- Under Hypothesis Tests, check **Exact** (in addition to Large-Sample).

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, click the Run button (the left-most button on the button bar at the top).

Confidence Limits Output

Confidence Intervals of Difference (P1-P2)

Confidence Interval Method	Estimated Value	Lower 95% Confidence Limit	Upper 95% Confidence Limit
Exact	0.2876	0.0177	0.5258
Quasi-Exact (Chen)	0.2876	0.0208	0.5230
Score (Farrington & Manning)	0.2876	0.0255	0.5124
Score (Miettinen & Nurminen)	0.2876	0.0227	0.5145
Score w/Skewness (Gart-Nam)	0.2876	0.0268	0.5196
Score (Wilson)	0.2876	0.0245	0.4989
Score (Wilson C.C.)	0.2876	-0.0044	0.5200
Chi-Square C.C. (Yates)	0.2876	0.0003	0.5750
Chi-Square (Pearson)	0.2876	0.0412	0.5340

This report now includes two additional tests that are exact tests. These tests will not be calculated if the total sample size is greater than the limit set under the Reports tab.

Hypothesis Tests

Two-Sided Tests of Zero Difference (H0: P1 = P2 versus H1: P1 <> P2)
Estimated Difference (P1 - P2) = 0.2876

Test Name	Test Statistic's Distribution	Test Statistic Value	Prob Level	Conclude H1 at 5% Significance?
Fisher's Exact	Hypergeometric		0.0388	Yes
Cond. Exact (Fisher)	Hypergeometric	5.921	0.0388	Yes
Cond. Exact (Pearson)	Hypergeometric	4.591	0.0388	Yes
Cond. Exact (Lik. Ratio)	Hypergeometric	4.763	0.0388	Yes
Exact (Barnard)	Double Binomial	2.143	0.0350	Yes
Chi-Square Test	Chi-Square(1)	4.591	0.0321	Yes
Chi-Square Test (C.C.)	Chi-Square(1)	3.376	0.0661	No
Z-Test	Normal	2.143	0.0321	Yes
Z-Test (C.C.)	Normal	1.837	0.0661	No
Mantel-Haenszel Test	Normal	2.121	0.0339	Yes
Likelihood Ratio	Chi-Square(1)	4.763	0.0291	Yes
T-Test using 0's and 1's	Student's T(47)	1.963	0.0556	No

This report now includes both the large-sample and the exact tests.

Example 3 – Noninferiority Test

This section presents an example of a noninferiority test of the ratio. To run a noninferiority test, a lower bound must be set. In this example, the bound will be set at 0.80. That is, if the positive response rate of the experimental group is at least 80% of the control group, the experiment group is concluded to be non-inferior to the control group.

In this example, 14 of 26 receiving the standard treatment responded positively and 19 of 23 receiving the experimental treatment responded positively.

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Proportions – Two Independent window.

1 Open the Proportions – Two Independent procedure.

- On the menus, select **Analysis**, then **Proportions**, then **Proportions – Two Independent**. The Proportions – Two Independent procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Proportions – Two Independent window, select the **Data tab**.
- In the **A Count** box, enter **19**.
- In the **B Count** box, enter **4**.
- In the **C Count** box, enter **14**.
- In the **D Count** box, enter **12**.
- Under Statistic(s), check only the **Ratio** box.
- Under Confidence Intervals, uncheck all boxes.
- Under Hypothesis Tests, check **Large-Sample** and **Exact**.
- Under Hypothesis Tests, uncheck **H0=0**.
- Under Hypothesis Tests, check **H0=User-Specified Value**.
- Under Hypothesis Tests, uncheck **Two-Sided**.
- Under Hypothesis Tests, check **Upper One-Sided**.

3 Set the H0 values.

- Select the **Hypothesis Tests tab**.
- Set the **Ratio** to **0.8**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Noninferiority Test Output

Upper-Tail, One-Sided Tests of Ratio (H0: P1 / P2 = 0.800 versus H1: P1 / P2 > 0.800)
Estimated Ratio (P1 / P2) = 1.534

Test Name	Test Statistic's Distribution	Test Statistic Value	Prob Level	Conclude H1 at 5% Significance?
Exact (Miettinen & Nurminen)	Double Binomial	3.104	0.0010	Yes
Score (Farrington & Manning)	Normal	3.136	0.0009	Yes
Score (Miettinen & Nurminen)	Normal	3.104	0.0010	Yes
Score (Gart & Nam)	Normal	3.130	0.0009	Yes

This report gives both an exact and a large-sample noninferiority test. Note that in this case, all tests reject the null hypothesis and conclude the alternative hypothesis that P1 (the response rate in the treatment group) is at least 80% of P2 (the response rate in control group).

Example 4 – Equivalence Test

This section presents an example of an equivalence test of the ratio. To run an equivalence test, both upper and lower bounds of equivalence must be set. In this example, the lower bound will be set at 0.80 and the upper bound is set to 1.20. That is, if the positive response rate of the experimental group is at least 80% of the control group, the experiment group is concluded to be non-inferior to the control group.

In this example, 14 of 26 receiving the standard treatment responded positively and 19 of 23 receiving the experimental treatment responded positively.

You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Proportions – Two Independent window.

1 Open the Proportions – Two Independent window.

- On the menus, select **Analysis**, then **Proportions**, then **Proportions – Two Independent**. The Proportions – Two Independent procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Proportions – Two Independent window, select the **Data tab**.
- In the **A Count** box, enter **19**.
- In the **B Count** box, enter **4**.
- In the **C Count** box, enter **14**.
- In the **D Count** box, enter **12**.
- Under Statistic(s), check only the **Ratio** box.

- Under Confidence Intervals, uncheck all boxes.
- Under Hypothesis Tests, check **Large-Sample Tests** and **Exact Tests**.
- Under Hypothesis Tests, uncheck **H0=0**.
- Under Hypothesis Tests, check **H0=User-Specified Value**.
- Under Hypothesis Tests, uncheck **Two-Sided**.
- Under Hypothesis Tests, check **Equivalence**.

3 Set the H0 Value.

- Select the **Hypothesis Tests tab**.
- Under Equivalence Bounds, set the **Lower Bound for Equivalence of a Ratio** to **0.8**.
- Under Equivalence Bounds, set the **Upper Bound for Equivalence of a Ratio** to **1.2**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Equivalence Test Output

Equivalence Tests of Ratio (H0: P1 / P2 < 0.80 or P1 / P2 > 1.20 versus H1: Equivalence)
Estimated Ratio (P1 / P2) = 1.534

Test Name	Lower Test Statistic's Value	Lower Test Statistic's Prob	Upper Test Statistic's Value	Upper Test Statistic's Prob	Prob Level	Conclude H1 at 5% Significance?
Exact (M. N.)	3.104	0.0010	0.000	1.0000	1.0000	No
Score (F. M.)	3.136	0.0009	1.245	0.8934	0.8934	No
Score (M. N.)	3.104	0.0010	1.232	0.8911	0.8911	No
Score (G. N.)	3.130	0.0009	1.246	0.8936	0.8936	No

This report gives both an exact and three large-sample equivalence tests.

Test Name

This column gives the name of the test. The abbreviations are M. N. for Mittinen and Nurminen, F. M. for Farrington and Manning, and G. N. for Gart and Nam.

Lower Test Statistic's Value

The equivalence test is based on two, one-sided tests (TOST). This is the test statistic for the lower test.

Lower Test Statistic's Probability

The equivalence test is based on two, one-sided tests (TOST). This is the significance level for the lower test.

Upper Test Statistic's Value

The equivalence test is based on two, one-sided tests (TOST). This is the test statistic for the upper test.

Upper Test Statistic's Probability

The equivalence test is based on two, one-sided tests (TOST). This is the significance level for the upper test.

Prob Level

This is the significance level of the test. This value is the maximum of the lower and upper probabilities. If this value is less that 0.05, the null hypothesis of non-equivalence is rejected and equivalence is concluded.

Conclude H1 at 5% Significance?

If this value is ‘No’, equivalence is not established. If this value is ‘Yes’, equivalence is established.

Chapter 520

Two Correlated Proportions (McNemar)

Introduction

This module computes confidence intervals and hypothesis tests for quantities derived from two paired (correlated) proportions. These quantities include the difference, the (risk) ratio, and the odds ratio.

Historically, McNemar's test has been used to test the hypothesis that the two proportions are equal. Recently, more interest has been placed on obtaining confidence intervals than on a specific hypothesis test.

Experimental Design

A common design that uses this analysis is when two dichotomous responses are made on each subject. For example, each subject is measured once after treatment A and again after the application of treatment B. The response is '1' if the event of interest occurs or '0' otherwise.

This analysis also applies to *matched pairs* data in which each *case* subject is matched with a similar subject from a *control* group.

Comparing Two Correlated Proportions

Suppose you have two dichotomous measurements Y_1 and Y_2 on each of *N* subjects (the 'subject' may be a pair of matched individuals). The proportions p_1 and p_2 represent the success probabilities. That is,

$$p_1 = \Pr(Y_1 = 1)$$

$$p_2 = \Pr(Y_2 = 1)$$

The corresponding failure proportions are given by $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$, respectively.

A sample of *N* subjects is selected and the two variables are measured. The data from this design can be summarized in the following 2-by-2 table:

	$Y_1 = 1$	$Y_1 = 0$	Total
$Y_2 = 1$	*a*	*c*	
$Y_2 = 0$	*b*	*d*	
Total			*n*

The proportions p_1 and p_2 are estimated from these data using the formulae

$$\hat{p}_1 = \frac{a+b}{n} \text{ and } \hat{p}_2 = \frac{a+c}{n}$$

Three quantities which allow these proportions to be compared are

Quantity	**Computation**
Difference	$\Delta = p_1 - p_2$
Risk Ratio	$\phi = p_1 / p_2$
Odds Ratio	$\psi = \frac{p_{21}}{p_{12}}$

Confidence Intervals

Several methods for computing confidence intervals for proportion difference, proportion ratio, and odds ratio have been proposed. We now show the methods that are available in ***NCSS***.

Difference

Four methods are available for computing a confidence interval of the difference between the two proportions $\Delta = p_1 - p_2$. The lower (*L*) and upper (*U*) limits of these intervals are computed as follows. Note that $z = |z_{\alpha/2}|$ is the appropriate percentile from the standard normal distribution.

Newcombe (1998) conducted a comparative evaluation of ten confidence interval methods. He recommended that the modified Wilson score method be used instead of the Pearson Chi-square or the Yate's Corrected Chi-square.

Nam's Score

For details, see Nam (1997) or Tango (1998). The lower limit is the solution of

$$L = \inf\left\{\Delta_0 : \frac{\hat{\Delta} - \Delta_0}{\tilde{\sigma}_{\Delta_0}} < z\right\}$$

and the upper limit is the solution of

$$U = \sup\left\{ \Delta_0 : \frac{\hat{\Delta} - \Delta_0}{\tilde{\sigma}_{\Delta_0}} > -z \right\}$$

where $\tilde{\sigma}_{\Delta_0}$ is given by

$$\tilde{\sigma}_{\Delta} = \frac{\tilde{p}_{21} + \tilde{p}_{12} - \Delta^2}{n}$$

$$\tilde{p}_{21} = \left\{ \frac{-e + \sqrt{e^2 - 8f}}{4} \right\}$$

$$\tilde{p}_{12} = \tilde{p}_{21} - \Delta$$

$$e = -\hat{\Delta}(1 - \Delta) - 2(\hat{p}_{21} + \Delta)$$

$$f = \Delta(1 + \Delta)\hat{p}_{21}$$

Wilson's Score as modified by Newcombe

For further details, see Newcombe (1998c), page 2639. This is Newcombe's method 10.

$$L = \hat{\Delta} - \delta$$

$$U = \hat{\Delta} + \varepsilon$$

where

$$\delta = \sqrt{f_2^2 - 2\hat{\phi} f_2 g_3 + g_3^2}$$

$$\varepsilon = \sqrt{g_2^2 - 2\hat{\phi} g_2 f_3 + f_3^2}$$

$$f_2 = \frac{(a+b)}{n} - l_2$$

$$g_2 = u_2 - \frac{(a+b)}{n}$$

$$f_3 = \frac{(a+c)}{n} - l_3$$

$$g_3 = u_3 - \frac{(a+c)}{n}$$

and l_2 and u_2 are the roots of

$$\left| x - \frac{a+b}{n} \right| = z\sqrt{\frac{x(1-x)}{n}}$$

and l_3 and u_3 are the roots of

$$\left|x - \frac{a+c}{n}\right| = z\sqrt{\frac{x(1-x)}{n}}$$

$$\hat{\phi}$$

$$\hat{\phi} = \begin{cases} \dfrac{\max(\text{ad - bc - n / 2},0)}{\sqrt{(\text{a + b})(c+d)(a+c)(b+d)}} & \text{if } ad > bc \\ \dfrac{\text{ad - bc}}{\sqrt{(\text{a + b})(c+d)(a+c)(b+d)}} & \text{otherwise} \end{cases}$$

Note that if the denominator of $\hat{\phi}$ is zero, $\hat{\phi}$ is set to zero.

Asymptotic Wald Method

For further details, see Newcombe (1998c), page 2638.

$$L = \hat{\Delta} - zs_W$$

$$U = \hat{\Delta} + zs_W$$

where

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_2 = (b-c)/n$$

$$s_W^2 = \frac{(a+d)(b+c)+4bc}{n^3}$$

Asymptotic Wald Method with Continuity Correction

For details, see Newcombe (1998c), page 2638.

$$L = \hat{\Delta} - zs_W - \frac{1}{n}$$

$$U = \hat{\Delta} + zs_W + \frac{1}{n}$$

Risk Ratio

Two methods are available for computing a confidence interval of the risk ratio $\phi = p_1 / p_2$. Note that $z = |z_{\alpha/2}|$ is the appropriate percentile from the standard normal distribution.

Nam and Blackwelder (2002) present two methods for computing confidence intervals for the risk ratio. These are presented here. Note that the score method is recommended.

Score (Nam and Blackwelder)

For details, see Nam and Blackwelder (2002), page 691. The lower limit is the solution of

$$z(\phi) = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z(\phi) = -|z_{\alpha/2}|$$

where

$$z(\phi) = \frac{\sqrt{n}(\hat{p}_1 - \phi\hat{p}_2)}{\sqrt{\phi(\tilde{p}_{12} + \tilde{p}_{21})}}$$

and

$$\tilde{p}_{12} = \frac{-\hat{p}_1 + \phi^2(\hat{p}_2 + 2\hat{p}_{12}) + \sqrt{(\hat{p}_1 - \phi\hat{p}_2)^2 + 4\phi^2\hat{p}_{12}\hat{p}_{12}}}{2\phi(\phi+1)}$$

$$\tilde{p}_{21} = \phi\tilde{p}_{12} - (\phi - 1)(1 - \hat{p}_{22})$$

Asymptotic Wald (Nam and Blackwelder)

For details, see Nam and Blackwelder (2002), page 692. The lower limit is the solution of

$$z_W(\phi) = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z_W(\phi) = -|z_{\alpha/2}|$$

where

$$z_W(\phi) = \frac{\sqrt{n}(\hat{p}_1 - \phi\hat{p}_2)}{\sqrt{\phi(\hat{p}_{12} + \hat{p}_{21})}}$$

Odds Ratio

Sahai and Khurshid (1995) present two methods for computing confidence intervals of the odds ratio $\psi = p_{21} / p_{12}$. Note that the maximum likelihood estimate of this is given by

$$\hat{\psi} = b / c$$

Exact Binomial

The lower limit is

$$\psi_L = \frac{b}{(c+1)F_{\alpha/2,2c+2,2b}}$$

and the upper limit

$$\psi_U = \frac{b+1}{cF_{\alpha/2,2b+2,2c}}$$

where F is the ordinate of the F distribution.

Maximum Likelihood

The lower limit is

$$\psi_L = \exp\left\{\ln\left(\hat{\psi}\right) - z_{\alpha/2}s_{\hat{\psi}}\right\}$$

and the upper limit

$$\psi_U = \exp\left\{\ln\left(\hat{\psi}\right) + z_{\alpha/2}s_{\hat{\psi}}\right\}$$

where

$$s_{\hat{\psi}} = \sqrt{\frac{1}{b} + \frac{1}{c}}$$

Hypothesis Tests

Difference

This module tests three statistical hypotheses about the difference in the two proportions:

1. H_0: $p_1 - p_2 = \Delta$ versus H_A: $p_1 - p_2 \neq \Delta$; this is a *two-tailed test*.
2. H_{0L}: $p_1 - p_2 \leq -\Delta$ versus H_{AL}: $p_1 - p_2 > -\Delta$; this is a *one-tailed test*.
3. H_{0U}: $p_1 - p_2 \geq \Delta$ versus H_{AU}: $p_1 - p_2 < \Delta$; this is a *one-tailed test*.

McNemar Test

Fleiss (1981) presents a test that is attributed to McNemar for testing the two-tailed null hypothesis. This is calculated as

$$\chi_1^2 = \frac{(b-c)^2}{b+c}$$

McNemar Test with Continuity Correction

Fleiss (1981) also presents a continuity-corrected version of McNemar test. This is calculated as

$$\chi_1^2 = \frac{(|b-c|-1)^2}{b+c}$$

Wald Test

Lui et al. (2002) present a pair of large-sample, Wald-type *z* tests for testing the two one-tailed hypothesis about the difference $p_1 - p_2 = \Delta$. These are calculated as

$$z_L = \frac{\hat{\Delta} + \Delta - \frac{1}{2n}}{\hat{\sigma}} \text{ and } z_U = \frac{\hat{\Delta} - \Delta + \frac{1}{2n}}{\hat{\sigma}}$$

where

$$\hat{\sigma}^2 = \frac{\hat{p}_{21} + \hat{p}_{12} - \hat{\Delta}^2}{n}$$

$$\hat{\Delta} = \hat{p}_1 - \hat{p}_2$$

Nam Test

Lui et al. (2002) recommend a likelihood score test which was originally proposed by Nam (1997). Lui et al. (2002) recommend this test. The tests are calculated as

$$z_L = \frac{\hat{\Delta} + \Delta}{\tilde{\sigma}_L} \text{ and } z_U = \frac{\hat{\Delta} - \Delta}{\tilde{\sigma}_U}$$

where

$$\tilde{\sigma}_L = \sigma_{-\Delta}$$

$$\tilde{\sigma}_U = \sigma_{\Delta}$$

and

$$\sigma_D = \frac{\tilde{p}_{21} + \tilde{p}_{12} - D^2}{n}$$

$$\tilde{p}_{21} = \left\{ \frac{-e + \sqrt{e^2 - 8f}}{4} \right\}$$

$$\tilde{p}_{12} = \tilde{p}_{21} - D$$

$$e = -\hat{\Delta}(1 - D) - 2(\hat{p}_{21} + D)$$

$$f = D(1 + D)\hat{p}_{21}$$

Ratio

This module tests three statistical hypotheses about the difference in the two proportions:

1. H_0: $p_1 / p_2 = \phi$ versus H_A: $p_1 / p_2 \neq \phi$; this is a *two-tailed test*.
2. H_{0L}: $p_1 / p_2 \leq 1/\phi$ versus H_{AL}: $p_1 / p_2 > 1/\phi$; this is a *one-tailed test*.
3. H_{0U}: $p_1 / p_2 \geq \phi$ versus H_{AU}: $p_1 / p_2 < \phi$; this is a *one-tailed test*.

Nam Test

For details, see Nam and Blackwelder (2002), page 691. The test statistic for testing a specific value of ϕ is

$$z(\phi) = \frac{\sqrt{n}\left(\hat{p}_1 - \phi\hat{p}_2\right)}{\sqrt{\phi\left(\tilde{p}_{12} + \tilde{p}_{21}\right)}}$$

where

$$\tilde{p}_{12} = \frac{-\hat{p}_1 + \phi^2\left(\hat{p}_2 + 2\hat{p}_{12}\right) + \sqrt{\left(\hat{p}_1 - \phi\hat{p}_2\right)^2 + 4\phi^2\hat{p}_{12}\hat{p}_{12}}}{2\phi\left(\phi + 1\right)}$$

$$\tilde{p}_{21} = \phi\tilde{p}_{12} - \left(\phi - 1\right)\left(1 - \hat{p}_{22}\right)$$

Equivalence Tests

Equivalence tests are hypothesis tests in which the alternative hypothesis, not the null hypothesis, is equality. The test is set up so that when the null hypothesis is rejected, equality is concluded. This is just the opposite of the usual statistical hypothesis. For example, suppose an accurate diagnostic test has serious side effects, so a replacement test is sought. In this case, we are not interested in showing that the two tests are different, but rather that they are the same.

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 tests are equivalent—that is, their accuracy is about the same. This requires a two-sided hypothesis test. On the other hand, *noninferiority tests* are used when we want to show that a new (experimental) test is no worse than the existing (reference or gold-standard) test. This requires a one-sided hypothesis test.

In the discussion to follow, two ways of expressing difference are considered: the difference and the ratio. The simple difference between two proportions is perhaps the most straight forward way of expressing that these two proportions are difference. A difference of zero means that the two proportions are equal. Unfortunately, this method does not work well near zero or one. For example, suppose a diagnostic test achieves 95% accuracy and we wish to establish that a new test is within 7 percentage points of the original test. The acceptable range is from 95% - 7% = 88% to 95% + 7% = 102%. Of course, 102% is impossible.

A second method of setting up the hypotheses that does not suffer from this problem is to consider the ratio of the two proportions. A ratio of one indicates that the two proportions are equal. Using the ratio to define the hypotheses will not result in impossible values.

Equivalence Based on the Difference

The equivalence between two proportions may be tested using the following hypothesis which is based on the difference:

$$H_0\colon p_1 - p_2 \le -\Delta \quad or \quad H_0\colon p_1 - p_2 \ge \Delta \text{ versus } H_A\colon -\Delta < p_1 - p_2 < \Delta.$$

where Δ is an established equivalence limit.

This hypothesis is often tested at the alpha significance level by determining if $100(1-2\alpha)\%$ confidence limits are both between $-\Delta$ and Δ. A second method to test this hypothesis is to separate it into two sets of one-sided hypotheses called the *non-inferiority hypothesis* and the *non-superiority hypothesis*.

Non-Inferiority Test

$H_{0L}: p_1 - p_2 \le -\Delta$ versus $H_{AL}: p_1 - p_2 > -\Delta$.

Non-Superiority Test

$H_{0U}: p_1 - p_2 \ge \Delta$ versus $H_{AU}: p_1 - p_2 < \Delta$.

Nam Test

Both the two, one-sided hypotheses method and the confidence interval test are tested using the score test proposed by Nam (1997). These tests and the confidence interval where presented above.

Ratio

The equivalence between two proportions may be tested using the following hypothesis which is based on the ratio:

$$H_0: p_1 / p_2 \le 1/\phi \quad or \quad H_0: p_1 / p_2 \ge \phi \text{ versus } H_A: 1/\phi < p_1 / p_2 < \phi.$$

where ϕ is an established equivalence limit which is assumed to be, without loss of generality, greater than one.

This hypothesis is often tested at the alpha significance level by determining if $100(1-2\alpha)\%$ confidence limits are both between $1/\phi$ and ϕ. A second method to test this hypothesis is to separate it into two sets of one-sided hypotheses called the *non-inferiority hypothesis* and the *non-superiority hypothesis*.

Non-Inferiority Test

$H_{0L}: p_1 / p_2 \le 1/\phi$ versus $H_{AL}: p_1 / p_2 > 1/\phi$.

Non-Superiority Test

$H_{0U}: p_1 / p_2 \ge \phi$ versus $H_{AU}: p_1 / p_2 < \phi$.

Nam Test

Both the two, one-sided hypotheses method and the confidence interval test are tested using the score test proposed by Nam and Blackwelder (2002). These tests and the confidence interval where presented above.

Procedure Options

This section describes the options available in this procedure.

Data Tab

Enter the data values directly on this panel.

Data Values

A Count

This is the number of individuals the responded positively (with a ‘Yes’) on both variables. This must be a non-negative number.

B Count

This is the number of individuals the responded positively (with a ‘Yes’) on the first variable and negatively (with a ‘No’) on the second variable. This must be a non-negative number.

C Count

This is the number of individuals the responded negatively (with a ‘No’) on the first variable and positively (with a ‘Yes’) on the second variable. This must be a non-negative number.

D Count

This is the number of individuals the responded negatively (with a ‘No’) on both variables. This must be a non-negative number.

Null Hypothesis Details

H0 Difference

This is the difference hypothesized by the null hypothesis. For a regular hypothesis test, this value will be zero. However, for equivalence and inferiority tests, this value will be nonzero.

Since this is the difference of two proportions, the range of values is from -1 to 1.

H0 Ratio

This is the ratio hypothesized by the null hypothesis. For a regular hypothesis test, this value will be one. However, for equivalence and inferiority tests, this value will be other than one.

Since this is the ratio of two proportions, the range of values is from 0 to infinity. Note that ‘0’ is not allowed.

Alpha

Confidence Limits

The quantity (1 - alpha) is the confidence coefficient of the confidence intervals. A 100 x (1 - alpha)% confidence interval will be calculated. This must be a value between 0.0 and 0.5.

Hypothesis Test

The probability in a hypothesis test of rejecting the null hypothesis (H0) when it is true. This is the specified significance level.

Equivalence Details

Max Equivalence Difference

This is the largest value of the difference that will still result in the conclusion of equivalence. Usually, this value is between 0 and 0.2.

Max Equivalence Ratio

This is the largest value of the ratio that will still result in the conclusion of equivalence. Usually, this value will be between 1.0 and 2.0.

Decimal Places

Proportions – Test Values

The number of digits displayed to the right of the decimal place.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Analysis of Two Correlated Proportions

This section presents an example of how to run an analysis on hypothetical data. In this example, two dichotomous variables where measured on each of fifty subjects; 30 subjects scored 'yes' on both variables, 9 subjects scored 'no' on both variables, 6 scored a 'yes' and then a 'no', and 5 scored a 'no and then a 'yes'.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Proportions – Two Correlated (McNemar) window.

1 Open the Proportions – Two Correlated window.

- On the menus, select **Analysis**, then **Proportions**, then **Proportions – Two Correlated (McNemar)**. The Proportions – Two Correlated (McNemar) procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Two Proportions window, select the **Data tab**.
- In the **A Count** box, enter **30**.
- In the **B Count** box, enter **6**.
- In the **C Count** box, enter **5**.
- In the **D Count** box, enter **9**.
- Set the **H0 Difference** box to **0.0**.
- Set the **H0 Ratio** box to **1.0**.
- Set the **Alpha - Confidence Limits** box to **0.05**.
- Set the **Alpha - Hypothesis Test** box to **0.05**.
- Set the **Max Equivalence Difference** box to **0.1**.
- Set the **Max Equivalence Ratio** box to **1.1**.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data and Proportion Section

Data Section

Data Source	A Yes-Yes	B Yes-No	C No-Yes	D No-No	N Total
Input	30	6	5	9	50

Proportion Section

Variable	Proportion Yes's	Lower 95.0% Conf. Limit of Proportion	Upper 95.0% Conf. Limit of Proportion	Number of Yes's	Number of No's	Total
One	0.7200	0.5833	0.8253	36	14	50
Two	0.7000	0.5625	0.8090	35	15	50

The Data Section report shows the values that were input. The Proportion Section gives the proportion of Yes's for each variable along with a large sample confidence interval based on the Wilson score method.

Confidence Intervals for the Difference (P1-P2)

Confidence Interval Method	Lower 95.0% Confidence Limit	Estimated Value	Upper 95.0% Confidence Limit
Score (Nam RMLE)*	-0.1197	0.0200	0.1606
Score (Wilson)	-0.1145	0.0200	0.1537
Asymptotic Wald	-0.1099	0.0200	0.1499
Asymptotic Wald (C.C.)	-0.1299	0.0200	0.1699

This report gives four confidence intervals for the difference. The formulas were given earlier in the technical details section. The Nam RMLE confidence interval is recommended.

Confidence Intervals for the Ratio (P1/P2)

Confidence Interval Method	Lower 95.0% Confidence Limit	Estimated Value	Upper 95.0% Confidence Limit
Score (Nam Blackwelder)*	0.8396	1.0286	1.2668
Asymptotic Wald	0.8567	1.0286	1.2350

This report gives two confidence intervals for the ratio. The formulas were given earlier in the technical details section. The Nam Blackwelder confidence interval is recommended.

Confidence Intervals for the Odds Ratio (P12/P21)

Confidence Interval Method	Lower 95.0% Confidence Limit	Estimated Value	Upper 95.0% Confidence Limit
Exact Conditional Binomial	0.3051	1.2000	4.9706
Maximum Likelihood	0.3662	1.2000	3.9320

This report gives two confidence intervals for the odds ratio. The formulas were given earlier in the technical details section. Neither confidence interval is recommended in all situations.

Two-Sided Hypothesis Tests about the Difference (P1-P2)

Name	Distribution of Test Statistic	Null Hypothesis (H0)	Test Statistic Value	Prob Level	Conclusion at the 5.0% Level
Nam*	CS(1)	P1-P2=0	0.09	0.7630	Cannot reject H0
McNemar	CS(1)	P1-P2=0	0.09	0.7630	Cannot reject H0
McNemar C.C.	CS(1)	P1-P2=0	0.00	1.0000	Cannot reject H0
Wald	CS(1)	P1-P2=0	0.20	0.6508	Cannot reject H0

This report gives four hypotheses test for a specified value of the difference. The Nam test is recommended. Although four tests are given, you should pick and use only one of these tests before seeing the results.

The formulas were given earlier in the technical details section.

Name

The name (author) of the test. Note that you should only use one of these tests.

Distribution of Test Statistic

This is the distribution of the test statistic on which the test is based. In this case, all four test statistics are chi-squares with one degree of freedom.

Null Hypothesis (H0)

The null hypothesis is the hypothesis that you hope to reject. The alternative hypothesis (the opposite of the null hypothesis) is concluded when the null hypothesis is rejected.

Test Statistic Value

This is the value of the test statistic under the null hypothesis. This is the chi-square value.

Prob Level

This is the *p-value* of the test. It is also called the significance level.

Conclusion at the 5.0% Level

This is the conclusion that is reached. Note that the conclusion 'Cannot reject H0' does not mean that the null hypothesis is concluded to be correct. It simply means that there was not enough evidence in the data to reject it.

One-Sided Hypothesis Tests about the Difference (P1-P2)

Name	Distribution of Test Statistic	Null Hypothesis (H0)	Test Statistic Value	Prob Level	Conclusion at the 5.0% Level
Nam Lower*	Normal	P1-P2<=0	0.30	0.3815	Cannot Reject H0
Nam Upper*	Normal	P1-P2>=0	0.30	0.6185	Cannot reject H0
Wald Lower	Normal	P1-P2<=0	0.45	0.3254	Cannot Reject H0
Wald Upper	Normal	P1-P2>=0	0.45	0.6746	Cannot reject H0

This report gives two hypotheses test for a specified value of the difference. Although two tests are given, you should pick and use only one of these tests before seeing the results.

Note that a separate test is given for each null hypothesis. The formulas were given earlier in the technical details section. The Nam test is recommended.

Name

The name (author) of the test. Note that you should only use one of these tests.

Distribution of Test Statistic

This is the distribution of the test statistic on which the test is based. In this case, all test statistics are standard normals.

Null Hypothesis (H0)

The null hypothesis is the hypothesis that you hope to reject. The alternative hypothesis (the opposite of the null hypothesis) is concluded when the null hypothesis is rejected.

Test Statistic Value

This is the value of the test statistic under the null hypothesis. This is the *z* value.

Prob Level

This is the *p-value* of the test. It is also called the significance level.

Conclusion at the 5.0% Level

This is the conclusion that is reached. Note that the conclusion 'Cannot reject H0' does not mean that the null hypothesis is concluded to be correct. It simply means that there was not enough evidence in the data to reject it.

Two-Sided Hypothesis Tests about the Ratio (P1/P2)

Name	Distribution of Test Statistic	Null Hypothesis (H0)	Test Statistic Value	Prob Level	Conclusion at the 5.0% Level
Nam*	Normal	P1/P2=1	0.09	0.7630	Cannot reject H0

This report gives the results of a two-sided hypotheses test for a specified value of the ratio. The formulas were given earlier in the technical details section.

Name

This is the name (author) of the test.

Distribution of Test Statistic

This is the distribution of the test statistic on which the test is based.

Null Hypothesis (H0)

The null hypothesis is the hypothesis that you hope to reject. The alternative hypothesis (the opposite of the null hypothesis) is concluded when the null hypothesis is rejected.

Test Statistic Value

This is the value of the test statistic under the null hypothesis. In this case, the test statistic is a *z* value.

Prob Level

This is the *p-value* of the test. It is also called the significance level.

Conclusion at the 5.0% Level

This is the conclusion that is reached. Note that the conclusion 'Cannot reject H0' does not mean that the null hypothesis is concluded to be correct. It simply means that there was not enough evidence in the data to reject it.

One-Sided Hypothesis Tests about the Ratio (P1/P2)

Name	Distribution of Test Statistic	Null Hypothesis (H0)	Test Statistic Value	Prob Level	Conclusion at the 5.0% Level
Nam Lower*	Normal	P1-P2<=0	0.30	0.3815	Cannot Reject H0
Nam Upper*	Normal	P1-P2>=0	0.30	0.6185	Cannot reject H0
Wald Lower	Normal	P1-P2<=0	0.45	0.3254	Cannot Reject H0
Wald Upper	Normal	P1-P2>=0	0.45	0.6746	Cannot reject H0

This report gives two hypotheses tests for a specified value of the ratio. Note that a separate test is given for each possible null hypothesis. The formulas were given earlier in the technical details section. The Nam test is recommended.

Name

The name (author) of the test. Note that you should only use one of these tests.

Distribution of Test Statistic

The distribution of the test statistic on which the test is based. In this case, both test statistics are standard normal.

Null Hypothesis (H0)

The null hypothesis is the hypothesis that you hope to reject. The alternative hypothesis (the opposite of the null hypothesis) is concluded when the null hypothesis is rejected.

Test Statistic Value

This is the value of the test statistic under the null hypothesis. This is the *z* value.

Prob Level

This is the *p-value* of the test. It is also called the significance level.

Conclusion at the 5.0% Level

This is the conclusion that is reached. Note that the conclusion 'Cannot reject H0' does not mean that the null hypothesis is concluded to be correct. It simply means that there was not enough evidence in the data to reject it.

Tests of Equivalence using Nam Score Confidence Intervals

Parameter Tested	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Min. Equiv. Bound Attaining 5.0% Signif.	Conclusion at 5.0% Level
Difference (P1-P2)	0.1219	-0.0950	0.1359	-0.1000	0.1000	0.1359	Cannot Reject H0
Ratio (P1/P2)	0.2418	0.8717	1.2188	0.9091	1.1000	1.2188	Cannot Reject H0

This report gives the results of two equivalence tests: one for a specified difference and one for a specified ratio. The formulas were given earlier in the technical details section.

Parameter Tested

This is the parameter being tested.

Prob Level

This is the *p-value* of the test. It is also called the significance level. In this case, it is the maximum p-value of the corresponding two one-sided hypothesis tests.

Lower/Upper 90.0% Conf. Limit

These are the lower and upper limits of a $100(1-2\alpha)\%$ confidence interval for the parameter tested on this line of the report. Note that the null hypothesis of non-equivalence is rejected when these confidence limits are within the lower and upper equivalence bounds.

Min. Equiv. Bound Attaining 5.0% Signif.

This is the smallest equivalence bound that just attains significance with these data.

Conclusion at the 5.0% Level

This is the conclusion that is reached. Note that the conclusion 'Cannot reject H0' does not mean that the null hypothesis is concluded to be correct. It simply means that there was not enough evidence in the data to reject it.

Chapter 525

Mantel-Haenszel Test

Introduction

The Mantel-Haenszel test compares the odds ratios of several 2-by-2 tables. Each table is of the form:

	Disease		
Exposure	**Yes (Cases)**	**No (Controls)**	**Total**
Yes	A	B	m_1
No	C	D	m_2
Total	n_1	n_2	n

where A, B, C, and D are counts of individuals.

The odds of an exposed individual contracting the disease is:

$$o_1 = \frac{p_1}{1-p_1} = \frac{A/m_1}{B/m_1} = \frac{A}{B}$$

The odds of an unexposed individual contracting the disease is:

$$o_2 = \frac{p_2}{1-p_2} = \frac{C/m_2}{D/m_2} = \frac{C}{D}$$

The odds ratio, ψ, is calculated using the equation:

$$\psi = \frac{o_1}{o_2} = \frac{AD}{BC}$$

This is the ratio of the odds of exposed individuals contracting a disease to the odds of unexposed individuals contracting a disease. It is closely related to the risk ratio which is:

$$Risk = \frac{p_1}{p_2}$$

When the probability of the disease is rare, $1-p_1 \approx 1$ and $1-p_2 \approx 1$. Hence, in this case, the odds ratio is almost identical to the risk ratio.

The above 2-by-2 table may be partitioned according to one or more variables into several 2-by-2 tables. Each individual table is referred to as a stratum. For example, consider the following data presented by Schlesselman (1982) for a case-control study investigating the relationship among lung cancer (the disease variable), employment in shipbuilding (exposure to asbestos), and smoking:

Smoking	Shipbuilding	Cancer	Control	Odds Ratio
Minimal	Yes	11	35	
	No	50	203	1.28
Moderate	Yes	70	42	
	No	217	220	1.69
Heavy	Yes	14	3	
	No	96	50	2.43

These data are contained in the SMOKING database. We see that the odds ratios steadily increase as the amount of smoking increases.

The Mantel-Haenszel analysis provides two closely related pieces of information. First, it provides statistical tests of whether the odds ratios are equal (homogeneous) or unequal (heterogeneous) across strata. Second, it provides an estimate of the odds ratio of the exposure variable, adjusted for the strata variable. In this example, it provides estimates of the odds ratio of asbestos exposure to lung cancer after removing the influence of smoking.

Assumptions

Two basic assumptions should be considered when using this procedure.

1. *Observations are independent from each other*. In practice, this means that each observation comes from a different subject, that the subjects were randomly selected from the population of interest, and that no specific group of subjects is purposefully omitted.
2. *All observations are identically distributed*. This means that they are obtained in the same way. For example, you could not mix the results of a telephone survey with those of a door-to-door survey.

Data Structure

The data may be entered in either raw or summarized form. In either case, each variable represents a factor and each row of data represents a cell. An optional variable may be used to give the count (frequency) of the number of individuals in that cell. When the frequency variable is left blank, each row receives a frequency of one. The following table shows how the above table was entered in the SMOKING database.

SMOKING dataset

Smoking	Shipbuilding	Cancer	Count
1_Minimal	1_Yes	1_Yes	11
1_Minimal	2_No	1_Yes	50
1_Minimal	1_Yes	2_No	35
1_Minimal	2_No	2_No	203
2_Moderate	1_Yes	1_Yes	70
2_Moderate	2_No	1_Yes	217
2_Moderate	1_Yes	2_No	42
2_Moderate	2_No	2_No	220
3_Heavy	1_Yes	1_Yes	14
3_Heavy	2_No	1_Yes	96
3_Heavy	1_Yes	2_No	3
3_Heavy	2_No	2_No	50

You will note that we have used the phrase “1_Yes” instead of “Yes” and “2_No” instead of “No.” We have done this so that the categories are sorted in the correct order. Normally, a “Yes” would be sorted after a “No.” However, the data in the table must be arranged so that the upper-left cell (corresponding to the count *A*) refers to individuals who are exposed to the risk factor and have the disease. Entering “1_Yes” and “2_No” causes the categories to be sorted in the proper order. An alternative way of accomplishing this would have been to enter a “1” for Yes and a “2” for No.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Binary Response Variable

Disease Variable

This is the binary response variable. The lowest value represents having the response. The highest value represents individuals who do not have the response. In a case-control study, this is the variable that separates the cases from the controls. The cases must receive the lowest value when the two categories are sorted, while the controls must be associated with the highest value.

Count Variable

Count Variable

This optional variable contains the count (frequency) of each cell in the table. If this variable is left blank, the frequency or each row in the database is assumed to be one.

Exposure Variable

Exposure Variable

This is the binary variable representing the risk factor. The lowest value represents individuals who have the risk factor. The highest value represents individuals who do not have the risk factor.

Delta

Delta

This value is added to each cell count in 2-by-2 tables with zeros in one or more cells. The traditional value is 0.5. Recent simulation studies have indicated that 0.25 produces better results under certain situations. Specific statistics that are impacted are indicated later. The Mantel-Haenszel statistic, for example, does not use this delta value in its calculation.

Strata Specification

Stratum (1-4) Variable

At least one factor variable must be specified. Examples of factor variables are gender, age groups, “yes” or “no” responses, etc.

A factor’s categories need not be consecutive integers. However, the program will display results in sorted order.

You may use text or numeric identification codes. The treatment of text variables is specified for each variable by the Data Type option on the Variable Info sheet.

Report Options

Alpha

This option specifies the significance level (alpha) used in calculating confidence limits.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Value Labels

Indicate whether to display the data values or their labels.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Mantel-Haenszel Test

This section presents an example of how to run an analysis of the data contained in the SMOKING database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Mantel-Haenszel Test window.

1 Open the SMOKING dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Smoking.s0**.
- Click **Open**.

2 Open the Mantel-Haenszel Test window.

- On the menus, select **Analysis**, then **Proportions**, then **Mantel-Haenszel Test**. The Mantel-Haenszel Test procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Mantel-Haenszel Test window, select the **Variables tab**.
- Double-click in the **Disease Variable** text box. This will bring up the variable selection window.
- Select **Cancer** from the list of variables and then click **Ok**. "Cancer" will appear in the Disease Variables box.
- Double-click in the **Count Variable** text box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**. "Count" will appear in the Count Variable box.
- Double-click in the **Exposure Variable** text box. This will bring up the variable selection window.
- Select **Shipbuilding** from the list of variables and then click **Ok**. "Shipbuilding" will appear in the Exposure Variables box.
- Double-click in the **Stratum 1 Variable** text box. This will bring up the variable selection window.
- Select **Smoking** from the list of variables and then click **Ok**. "Smoking" will appear in the Stratum 1 Variables box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Strata Count Section

Strata Count Section

Strata	Smoking	A	B	C	D	Sample Odds Ratio
1	1_Minimal	11	35	50	203	1.2760
2	2_Moderate	70	42	217	220	1.6897
3	3_Heavy	14	3	96	50	2.4306

A: Shipbuilding = 1_Yes, Cancer = 1_Yes
B: Shipbuilding = 1_Yes, Cancer = 2_No
C: Shipbuilding = 2_No, Cancer = 1_Yes
D: Shipbuilding = 2_No, Cancer = 2_No

This report presents the data that were input. Each row of the report represents an individual 2-by-2 table. The definitions of the four letters (A, B, C, and D) are shown immediately below the report. Thus A is the number of individuals with Disease = Yes and Exposure = Yes.

Sample Odds Ratio

This is the odds ratio calculated for the 2-by-2 table listed on this row. The formula is

$$\psi = \frac{AD}{BC}$$

The sample odds ratio is not calculated when any of the four cell counts is zero. Note that this value is different from the Corrected Odds Ratio report in the Strata Detail Section.

Strata Detail Section

Strata Detail Section

Strata	Lower 95.0% C.L.	1/2-Corrected Odds Ratio	Upper 95.0% C.L.	Exact Test	Proportion Exposed	Proportion Diseased
1	0.5653	1.2909	2.8335	0.5516	0.1538	0.2040
2	1.0805	1.6857	2.6465	0.0194	0.2040	0.5228
3	0.6127	2.2891	11.2098	0.1870	0.1043	0.6748

Each line on this report presents results for an individual 2-by-2 table. The strata number provides the identity of particular 2-by-2 table, since the tables in this report are listed in the same order as those in the Strata Count Section report described previously.

Strata

The row number.

1/2-Corrected Odds Ratio

This odds ratio is computed using the formula:

$$\psi' = \frac{(A+\delta)(D+\delta)}{(B+\delta)(C+\delta)}$$

where δ is the Delta value that was entered (usually, 0.5 or 0.25). Note that this odds ratio is defined when one or more cell counts are zero.

Lower, Upper 100(1-Alpha)% C.L.

The odds ratio confidence limits are calculated from those based on the Log Odds Ratio using the following procedure.

1. Compute the corrected odds ratio ψ' using the formula above.

2. Compute the logarithm of the odds ratio using:

$$L' = \ln(\psi')$$

3. Compute the standard error of *L'* using:

$$s_{L'} = \sqrt{\frac{1}{(A+\delta)} + \frac{1}{(B+\delta)} + \frac{1}{(C+\delta)} + \frac{1}{(D+\delta)}}$$

4. Compute the *100(1-α)%* confidence limits for *L* using the fact that *L'* is approximately normally distributed for large samples:

$$L' \pm z_{\alpha/2} s_{L'}$$

   where $z_{\alpha/2}$ is the appropriate value from the standard normal distribution.

5. Transform the above confidence limits back to the original scale using:

$$\psi_{lower} = e^{L' - z_{\alpha/2} s_{L'}}$$

$$\psi_{upper} = e^{L' + z_{\alpha/2} s_{L'}}$$

6. Compute the quantities *X*, *Y*, *W*, *F*, *U*, and *V* using the formulas:

$$X = \psi(m_1 + n_1) + (m_2 - n_1)$$

$$Y = \sqrt{X^2 - 4m_1 n_1 \psi(\psi - 1)}$$

$$N_{11} = \frac{X - Y}{2(\psi - 1)}$$

$$N_{12} = m_1 - N_{11}$$

$$N_{21} = n_1 - N_{11}$$

$$N_{22} = n_2 - m_1 + N_{11}$$

$$W = \frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}$$

$$F = \left(A - N_{11} - \tfrac{1}{2}\right)^2 W - z_{\alpha/2}^2$$

$$T = \frac{1}{2(\psi - 1)^2}\left(Y - n - \frac{\psi - 1}{Y}\left[X(m_1 + n_1) - 2m_1 n_1(2\psi - 1)\right]\right)$$

$$U = \frac{1}{N_{12}^2} + \frac{1}{N_{21}^2} - \frac{1}{N_{11}^2} - \frac{1}{N_{22}^2}$$

$$V = T\left[\left(A - N_{11} - \tfrac{1}{2}\right)^2 U - 2W\left(A - N_{11} - \tfrac{1}{2}\right)\right]$$

Finally, use the updating equation below to calculate a new value for the odds ratio.

$$\psi^{(k+1)} = \psi^{(k)} - \frac{F}{V}$$

7. Continue iterating (computing the values in step 6) until the value of F is arbitrarily close to zero (say, if its absolute value is less than 0.0000001).

This procedure is used separately for the upper and lower confidence limits of the odds ratio.

Exact Test

This is the probability (significance) level of Fisher's exact test versus a two sided alternative. Reject the hypothesis that $\psi = 1$ when this value is less than a small value, say 0.05.

Proportion Exposed

This is the overall proportion of those in the table that were exposed to the risk factor. The calculation is:

$$m_1 / n$$

This figure may or may not estimate this proportion in the population depending on the sampling design that was used to obtain the data.

Proportion Diseased

This is the overall proportion of those in the table that were diseased. The calculation is:

$$n_1 / n$$

This figure may or may not estimate this proportion in the population depending on the sampling design that was used to obtain the data.

Mantel-Haenszel Statistics Section

Mantel-Haenszel Statistics Section

Method	Lower 95.0% C.L.	Estimated Odds Ratio	Upper 95.0% C.L.	Chi-Square Value	DF	Prob Level
MH C.C.	1.1411	1.6438	2.3679	7.12	1	0.007616
MH	1.1545	1.6438	2.3404	7.60	1	0.005832
Robins	1.1558	1.6438	2.3378			
Woolf	1.1417	1.6291	2.3247	7.24	1	0.007137
Heterogeneity Test				0.81	2	0.667190

This report presents various odds ratio confidence limits and hypothesis tests. From what has been written in the statistical literature, we would recommend the following strategy when using this report.

For hypothesis testing, use the Heterogeneity Test to test the hypothesis that all odds ratios are equal. Use the MH C.C. hypothesis test to test the hypothesis test that all odds ratios are equal to one. Use the Robins confidence limits.

We will next discuss each row of this report.

MH C.C.

This row presents the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. Generally speaking, the continuity correction is used to provide a closer approximation to the exact conditional test in which all marginal totals are assumed to be fixed.

Hypothesis Test

We will discuss the hypothesis test first. The Mantel-Haenszel chi-square value tests the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. Note that this is different from the Heterogeneity Test which does not test that the odds ratios are equal to one, just equal to each other.

The formula used for this test is:

$$\chi^2_{mhcc} = \frac{\left(\left|\sum_{i=1}^{K} A_i - \sum_{i=1}^{K} E(A_i)\right| - \frac{1}{2}\right)^2}{\sum_{i=1}^{K} V(A_i)}$$

where K is the number of strata and

$$E(A_i) = \frac{n_{1i} m_{1i}}{n_i}$$

$$V(A_i) = \frac{n_{1i} n_{2i} m_{1i} m_{2i}}{n_i^2 (n_i - 1)}$$

This is a chi-square test with one degree of freedom. The probability level provides the upper tail probability of the test. Hence, when this value is less that your desired alpha level (say 0.05), reject the null hypothesis that all odds ratios are equal to one.

Confidence Limits for the Odds Ratio

The within-strata odds ratio is computed as follows:

$$\psi_{mh} = \frac{\sum_{i=1}^{K} \frac{A_i D_i}{n_i}}{\sum_{i=1}^{K} \frac{B_i C_i}{n_i}}$$

The test-based confidence limits for a 100(1-alpha)% confidence interval are given by

$$\psi_{mhc,lower} = \exp\left(\left(1 - \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mhcc}}}\right) \ln(\psi_{mh})\right)$$

$$\psi_{mhc,upper} = \exp\left(\left(1 + \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mhcc}}}\right) \ln(\psi_{mh})\right)$$

MH

This row presents the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. Generally speaking, the uncorrected version of this test is used to provide a closer approximation to the unconditional chi-square test.

Hypothesis Test

We will discuss the hypothesis test first. The Mantel-Haenszel chi-square value tests the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. Note that this is different from the Heterogeneity Test which does not test that the odds ratios are equal to one, just equal to each other.

The formula used for this test is:

$$\chi^2_{mh} = \frac{\left(\left|\sum_{i=1}^{K} A_i - \sum_{i=1}^{K} E(A_i)\right|\right)^2}{\sum_{i=1}^{K} V(A_i)}$$

where K is the number of strata and

$$E(A_i) = \frac{n_{1i} m_{1i}}{n_i}$$

$$V(A_i) = \frac{n_{1i} n_{2i} m_{1i} m_{2i}}{n_i^2 (n_i - 1)}$$

This is a chi-square test with one degree of freedom. The probability level provides the upper tail probability of the test. Hence, when this value is less that your desired alpha level (say 0.05), reject the null hypothesis that all odds ratios are equal to one.

Confidence Limits for the Odds Ratio

The within-strata odds ratio is computed as follows:

$$\psi_{mh} = \frac{\sum_{i=1}^{K} \frac{A_i D_i}{n_i}}{\sum_{i=1}^{K} \frac{B_i C_i}{n_i}}$$

The test-based confidence limits for a 100(1-alpha)% confidence interval are given by

$$\psi_{mhc,lower} = \exp\left(\left(1 - \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mh}}}\right) \ln(\psi_{mh})\right)$$

$$\psi_{mhc,upper} = \exp\left(\left(1 + \frac{z_{\alpha/2}}{\sqrt{\chi^2_{mh}}}\right) \ln(\psi_{mh})\right)$$

Robins

Robins (1986) presented an alternative formulation for the confidence limits which they have shown to be more accurate that any of the others presented. They did not make modifications to the hypothesis test, so no hypothesis test is printed on this line of the report.

Confidence Limits for the Odds Ratio

The within-strata odds ratio is computed as follows:

$$\psi_R = \frac{\sum_{i=1}^{K} \frac{A_i D_i}{n_i}}{\sum_{i=1}^{K} \frac{B_i C_i}{n_i}}$$

The confidence limits for a 100(1-alpha)% confidence interval are given by

$$\psi_{R,lower} = \exp\left(\left(\ln\left(\psi_R\right) - z_{\alpha/2}\sqrt{Vus}\right)\right)$$

$$\psi_{R,upper} = \exp\left(\left(\ln\left(\psi_R\right) + z_{\alpha/2}\sqrt{Vus}\right)\right)$$

where

$$Vus = \frac{\sum_{i=1}^{K} P_i R_i}{2\left(\sum_{i=1}^{K} R_i\right)^2} + \frac{\sum_{i=1}^{K} P_i S_i + Q_i R_i}{2\left(\sum_{i=1}^{K} R_i\right)\left(\sum_{i=1}^{K} S_i\right)} + \frac{\sum_{i=1}^{K} Q_i S_i}{2\left(\sum_{i=1}^{K} S_i\right)^2}$$

$$P_i = \frac{A_i + D_i}{n_i}$$

$$Q_i = \frac{B_i + C_i}{n_i}$$

$$R_i = \frac{A_i D_i}{n_i}$$

$$S_i = \frac{B_i C_i}{n_i}$$

Woolf

This row presents the confidence limits and hypothesis test developed by Woolf, as described by Schlesselman (1982). Recent studies have cast doubt on the usefulness of Woolf's tests, but they are provided anyway for completeness.

Confidence Limits for the Odds Ratio

The within-strata odds ratio is computed as follows:

$$\psi_w = \exp\left(\frac{\sum_{i=1}^{K} v_i^{-1} \ln \psi_i}{\sum_{i=1}^{K} v_i^{-1}}\right)$$

where

$$v_i = \frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i}$$

If any of the counts in a particular 2-by-2 table are zero, all counts in that table have delta (0.25 or 0.5) added to them.

Let W be calculated by:

$$W = \sum_{i=1}^{K} v_i^{-1}$$

The confidence limits are calculate using the equations:

$$\psi_{w,lower} = \psi_w \exp\left(-\frac{z_{\alpha/2}}{\sqrt{W}}\right)$$

$$\psi_{w,upper} = \psi_w \exp\left(\frac{z_{\alpha/2}}{\sqrt{W}}\right)$$

Hypothesis Test

Woolf's chi-square statistic tests the hypothesis that all odds ratios are equal to one.

The formula used for this test is:

$$\chi_w^2 = W\left(\ln \psi_w\right)^2$$

This is a chi-square test with one degree of freedom. The probability level provides the upper tail probability of the test. Hence, when this value is less that your desired alpha level (say 0.05), reject the null hypothesis that all odds ratios are equal to one.

Heterogeneity Test

This row presents a hypothesis test developed by Woolf, as described by Schlesselman (1982) for testing the more general hypothesis that all odds ratios are equal, but not necessarily equal to one.

Hypothesis Test

Woolf's chi-square statistic tests the hypothesis that all odds ratios are equal.

The formula used for this test is:

$$\chi_{wh}^2 = \sum_{i=1}^{K} v_i^{-1}\left(\ln \psi_i - \ln \psi_w\right)^2$$

This is a chi-square test with K-1 degrees of freedom. The probability level provides the upper tail probability of the test. Hence, when this value is less that your desired alpha level (say 0.05), reject the null hypothesis that all odds ratios are equal to one.

Chapter 530

Loglinear Models

Introduction

Loglinear models (LLM) studies the relationships among two or more discrete variables. Often referred to as multiway frequency analysis, it is an extension of the familiar chi-square test for independence in two-way contingency tables.

LLM may be used to analyze surveys and questionnaires which have complex interrelationships among the questions. Although questionnaires are often analyzed by considering only two questions at a time, this ignores important three-way (and multi-way) relationships among the questions. The use of LLM on this type of data is analogous to the use of multiple regression rather that simple correlations on continuous data.

There are several textbooks available that explain LLM in detail. We recommend the books by Tabachnick (1989) and Jobson (1992) which each have excellent chapters on LLM. Wickens (1989) is a book that is completely devoted to LLM.

Limitations and Assumptions

Since the use of LLM requires few assumptions about population distributions, it is remarkably free of limitations. It may be applied to almost any circumstance in which the variables are (or can be made) discrete. It can even be used to analyze continuous variables which fail to meet distributional assumptions (by collapsing the continuous variables into a few categories).

Three basic assumptions should be considered when using LLM.

1. *Observations are independent from each other*. In practice, this means that each observation comes from a different subject, that the subjects were randomly selected from the population of interest, and that no specific group of subjects is purposefully omitted.
2. *All observations are identically distributed*. This means that they are obtained in the same way. For example, you could not mix the results of a telephone survey with those of a door-to-door survey.
3. *The number of observations is large*. Since LLM makes use of large sample approximations, it requires large samples. The LLM algorithm begins by taking the natural logarithm of each of the cell frequencies, so empty cells (those with frequencies of zero) are not allowed. LLM appears to be less restrictive than traditional chi-square contingency tests, so rules that are used for those tests may be used for LLM analysis as well.

Fundamental Approach

LLM analysis requires two steps. It is easy to become lost in details of each of these steps, but it is important to keep in mind the overall purpose of each task.

1. *Selecting an appropriate model*. The first step is to find an appropriate model of the data. Several techniques may be used to find an appropriate LLM. One of the most popular is the step-down technique in which complex terms are removed until all terms remaining are significant.

   This search for an appropriate model is restricted to those models which are *hierarchical*. Hierarchical models are those in which the inclusion of a term forces the inclusion of all components of that term. For example, the inclusion of the two-way interaction, AB, forces terms A and B to also be included.

   Before the model is accepted, you should study the residuals to determine if the model fits the data reasonably well.

2. *Interpreting the selected model*. Once a model is selected, it must be interpreted. This is the step in which you determine what your data are telling you.

The Notation of Loglinear Models

Consider a two-way table in which the row-variable ***A*** has categories (levels) *i=1,...,I* and the column-variable ***B*** has categories *j=1,...,J*. A multiplicative model that reproduces the cell frequencies f_{ij} exactly is

$$m_{ij} = N\alpha_i\beta_j\gamma_{ij}$$

where $m_{ij} = E\left(f_{ij}\right)$ is the expected frequency of the i^{th} row and the j^{th} column. When the m_{ij} are estimated using maximum likelihood, the results are denoted $\hat{m}_{ij}$. Also note that $N = \sum_{ij} f_{ij}$.

One aspect of the table that is of interest is whether ***A*** and ***B*** are independent. This is often tested using the familiar chi-square test. In the above formula, independence would be established if all γ_{ij} were equal to one.

Because of its multiplicative form, the above formula is difficult to work with. However, if we take the logarithm of both sides, we can rewrite it as

$$\ln\left(m_{ij}\right) = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

The $\lambda's$ are called *effects*. The superscript indicates the variable(s) and the subscripts refer to the individual categories of those variables. The *order* of an effect is equal to the number of variables in the superscript.

Because this formulation is additive, it is called a *loglinear* model. Because of the logarithms, this model has the added constraint that none of the m_{ij} are zero.

Notice that the total number of $\lambda's$ in this model is *1+I+J+(IxJ)* which is greater than the number of cell frequencies (which is *IxJ*). When the number of parameters is greater than or equal

to the number of cells, we say the model is *saturated*. A saturated model reproduces the observed frequencies exactly.

By testing whether certain of the $\lambda's$ are zero, you can test various interrelationships. For example, to test whether all of the frequencies are equal, you would test whether all first-order and second-order effects (the $\lambda^{A}'s, \lambda^{B}'s$, and $\lambda^{AB}'s$) are zero. Testing whether the $\lambda^{AB}'s$ are zero would test whether variables ***A*** and ***B*** are independent. Testing whether the $\lambda^{A}'s$ were zero would test whether the probabilities of the categories of ***A*** are equal. As you can see, this model will let you answer many interesting questions about factors ***A*** and ***B***.

Hierarchical Models

The three-way LLM would be written as

$$\ln\left(m_{ijk}\right) = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

or, using the familiar ANOVA syntax, it might be written as

$$y_{ijk} = mean + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk}$$

Various models that are subsets of this saturated model might be of interest. For example, the main-effects model, *A+B+C*, would be useful in testing whether the factors are independent.

Hierarchical models are a particular class of models in which no interaction term is specified unless all subset combinations of that term are also in the model.

Often, a shorthand notation is used to express these models in which only the largest terms are specified. The following examples, showing the hierarchical model on the left and the expanded model on the right, should give you the idea of how this notation works.

Hierarchical Notation	Regular (Expanded) Notation
ABC	A+B+AB+C+AC+BC+ABC
A,BC	A+B+C+BC
AB	A+B+AB
AB,BC	A+B+AB+C+BC
A,B	A+B
A,B,C	A+B+C
AB,AC,BC	A+B+AB+C+AC+BC

In the LLM analysis considered in this program, only hierarchical models are used. Hence, we adopt the shorthand model specification on the left, although we are actually fitting the expanded model on the right.

Goodness of Fit

When dealing with several competing models, the relative quality of each model must be considered. The quality of a model, as measured by its goodness of fit to the data, may be tested using either of two chi-square statistics:

The Pearson chi-square statistic

$$\chi^2 = 2\sum_{i,j,k} \frac{\left(f_{ijk} - \hat{m}_{ijk}\right)^2}{\hat{m}_{ijk}}$$

and the likelihood-ratio statistic

$$G^2 = 2\sum_{i,j,k} f_{ijk} \ln\left(\frac{f_{ijk}}{\hat{m}_{ijk}}\right)$$

Both of these statistics are distributed as a chi-square random variable when *N* is large and none of the $\hat{m}_{ijk}$ are small. If a few of the $\hat{m}_{ijk}$ are small, the chi-square approximation is still fairly close. Both of these statistics have *n-p* degrees of freedom where *n* is the number of cells in the table and *p* is the number of parameters in the model on which the $\hat{m}_{ijk}$ are based.

You should understand exactly what these two chi-square statistics are testing. They test whether the terms in the saturated model that are not included in the current model are significantly different from zero.

For example suppose the hierarchical model *AB, BC* is fit. The expanded version of this hierarchical model is *A+B+C+AB+BC*. Note that the terms *AC* and *ABC* are omitted. If the χ^2 and G^2 were computed using the $\hat{m}_{ijk}$ from this fit, they would test whether the *AC* and *ABC* effects are zero. That is, these chi-square statistics test whether any important effects have been left out of the model.

The likelihood-ratio statistic, G^2, enjoys a very useful property which the Pearson χ^2 does not have. It is additive under partitioning for nested models. To explain this, consider an example. Suppose the model *AB,AC,BC* is fit and the resulting value of G^2 is 17.8 with 8 degrees of freedom. A second model *A,B,C* is fit resulting in a G^2 value of 69.9 with 24 degrees of freedom. If you expand each of these models, you will find that the terms *AB*, *AC*, and *BC* are in the first model but not in the second. Also, note that the second model is nested (completely contained) in the first model. If you subtract the first G^2 from the second, you will get 52.1. This is also a valid chi-square statistic with degrees of freedom 24 - 8 = 16. It tests whether *AB, AC,* and *BC* are significant.

This additivity property is a very useful. It allows you to test the importance of various individual terms. For example, suppose the model *AB,AC,BC* is tested and the goodness-of-fit test is not significant. This means that this particular model, *A+B+C+AB+AC+BC*, fits the data adequately. The next question is whether all six of these terms are necessary. To test the significance of *BC* you would fit the model, *A+B+C+AB+AC*, and subtract the first G^2 value from the second. This would test the significance of *BC*.

A word of caution: the difference between the two G^2 is distributed as a chi-square only when the more complete model fits the data adequately. That means that the G^2 of the larger model should be nonsignificant. Because of the additivity property of G^2, it is very popular in LLM.

Again, this additivity property does not hold for the Pearson chi-square statistic. Why do we even compute this value? Why not just use the likelihood ratio statistic? For two reasons. First, some studies indicate that the Pearson goodness of fit test may be more accurate. Second, since both of these are asymptotic tests, you can be more comfortable with small sample results when both tests lead to the same conclusion.

Model Selection Techniques

One of the main tasks in working with LLM's is dealing with the large number of possible models that can be generated from a single data table. The number of terms in the saturated model doubles with each additional factor. For example, there are 16 effects in a four-factor study and 32 effects in a five-factor study. When you consider the number of possible models that can be created from the 16 effects in a four-factor study, you begin to see the magnitude of the task. Even limiting your search to just the hierarchical models still leaves you with a large number of models to consider. There are over 100 different hierarchical models in a four-factor study, and over a 1000 in a five-factor study.

Since your first task in the analysis is to find a well-fitting model with as few terms as possible, you must adopt some method to limit the number of models you consider. The program provides several possible model selection methods. The final model will result from applying several of these techniques to your data.

Standardized Parameter Estimates

This method screens models as follows. First, a standardized estimate of each λ in the saturated model is calculated. Next, a list is made of the largest effects (greater than some cutoff value like 2.0 or 3.0). Finally, a hierarchical model is selected which includes as few terms as possible while still including the list of significant effects. This model is tested for adequacy using the chi-square test. If the goodness-of-fit test is nonsignificant, the model is used. Otherwise, additional effects are added to the model (based on their standardized values) until an adequate model is found.

Tests of Marginal and Partial Association

This method computes two tests for each term (up to fourth order terms). These tests assume that terms of higher order are negligible. The two tests are for *partial* and *marginal* association. The partial association considers the significance of a term after considering all other terms of the same order. The marginal association tests the significance of the term ignoring the influence of the other factors in the model.

The *partial association* test is constructed as follows. Fit a model containing all terms with the same order as the term being tested. Fit a second model identical with the first except with the term of interest. Subtract the first likelihood-ratio statistic from the second. The degrees of freedom are also determined by subtraction.

For example, to test that the partial association between ***A*** and ***B*** is zero in a four-way table, compute the values of G^2 for the models *AB,AC,AD,BC,BD,CD* and *AC,AD,BC,BD,CD*. The difference between these two values tests the partial association.

The *marginal association* test is constructed by collapsing the table until the term of interest is the highest-order interaction and there are no other terms of the same order. This term is then removed and the next lowest model is fit. The G^2 value tests the marginal association among the factors in the term.

For example, to test that the marginal association between ***A*** and ***B*** is zero in a four-way table, first collapse the table to the two-way table containing only ***A*** and ***B***. Next fit the model *A,B* on the collapsed table and compute the value of G^2 . This G^2 value tests the marginal association between **A** and **B**.

By considering the results of these two tests for each term, you can gain a fairly good indication of which terms are significant and which are not. As before, to obtain the final model, make a list of all terms that are significant. Next, write down the minimal hierarchical model that includes these terms.

Simultaneous Order Tests

The program produces a report that simultaneously tests all terms of a given order and all terms of a given order and higher. These tests let you immediately reduce the number of models that must be considered. For example, if the test of second-order models and higher is significant while the test for third-order models and higher is not, you know that the maximum order that must be considered is two. This knowledge allows you to reduce your search to second-order models.

Step-Down Selection Procedure

This is probably the most popular model selection method. It is the method that is used by default in this program. This procedure begins with a specified model (often the saturated model is used since it fits the data well) and searches for a model with fewer terms that still fits well. The program uses a backward elimination selection technique, which works better than the forward selection technique.

This procedure works as follows. First, a significance level (alpha) is chosen for the goodness of fit test to signal a significant model (a model that does not fit the data). Next, each of the highest-order hierarchical terms is removed, being replaced with appropriate terms so that the resulting expanded model is only different by the term of interest. The G^2 values of the original model and the subset model are then differenced so that the term may be tested individually. The model picked is the sub-model having the largest significance probability. The procedure terminates when no sub-model can be found with a probability greater than alpha.

Analyzing the Residuals

Once a candidate model has been found, it must be further analyzed for adequacy. In addition to checking goodness of fit statistics, the residuals between the estimated and actual frequencies should be studied. If a particular cell seems to be causing distortion in the results, appropriate action must be taken (such as adding deleted terms back into the model).

Once the residuals appear to be okay, the various terms in the model must be interpreted. This is accomplished by considering the percentages in the corresponding collapsed tables.

Data Structure

The data may be entered in either raw or summarized form. In either case, each variable represents a factor and each row of data represents a cell. An optional variable may be used to give the frequency (count) of the number of individuals in that cell. When the frequency variable is left blank, each row receives a frequency of one.

The following data are a portion of the results of a study by Dyke and Patterson (1952) on the information sources people use to obtain their knowledge of cancer. The data are contained in the LOGLIN1 database.

LOGLIN1 dataset (subset)

Counts	Newspaper	Lecture	Radio	Reading	Knowledge
23	1	1	1	1	1
102	1	2	1	1	1
1	2	1	1	1	1
16	2	2	1	1	1
8	1	1	1	1	2
67	1	2	1	1	2
3	2	1	1	1	2
16	2	2	1	1	2
8	1	1	1	2	1
35	1	2	1	2	1

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Factor Specification

Factor Variables (A-G)

At least two factor variables must be specified. Examples of factor variables are gender, age groups, “yes” or “no” responses, etc.

The factor categories need not be consecutive integers. You may use text or numeric identification codes. The treatment of text variables is specified for each variable by the Data Type option on the Variable Info sheet.

The first variable listed becomes factor A, the second becomes factor B, and so on. Up to seven factor variables may be designated.

Frequency Variable

Frequency Variable

This optional variable contains the count (frequency) of each cell in the table. If this variable is left blank, the frequency or each row in the database is assumed to be one.

Delta

Delta Value

This value is added to each cell count. It is used to add a small amount (between 0.1 and 0.9) to each cell count when zeros are present in the table. Remember that since the algorithm begins by taking the logarithm of each cell frequency and since the logarithm of zero is not defined, you cannot analyze a table with zero counts. This option lets you analyze data with zero frequencies.

When using this option, consider running your analysis over with two or three difference delta values to determine if the delta value is making a difference in the outcome (it should not).

Model Specification

Model

This option allows you to specify the hierarchical model to be fit. If a step-down selection is to be run, this model will serve as the starting point.

- **Full Model**

  This option causes the saturated model to be fit.

- **Up to (1,2,3,4)-Way**

  These options indicate that only terms up to and including that order are kept in the model.

  For example, if you are studying four factors and you specify "2Way," you would analyze the hierarchical model *AB,AC,AD,BC,BD,CD*.

- **Custom**

  Selecting this option causes the model specified in the Custom Model option (which is next) to be used.

Custom Model

This box specifies a hierarchical model of your choice. The syntax of hierarchical models is as follows:

Hierarchical terms refer to many actual terms. For example, suppose you have six factors (labeled A-F) and you specify the hierarchical model: *ABC,BCD,DE,F*. The expanded model would be:

Hierarchical Term	Expanded Terms
ABC	A,B,AB,C,AC,BC,ABC
BCD	B,C,BC,D,BD,CD,BCD
DE	D,E,DE
F	F

Notice that the simple terms B, C, and BC are contained in both the ABC list and the BCD list. Also, the term D is contained in both the BCD list and the DE list. The actual model fit would be the complete list, with each multiple-entry only listed once. Here, the expanded model would be:

A,B,AB,C,AC,BC,ABC,D,CD,BD,BCD,E,DE,F

If you want to specify a saturated model, you simply enter the highest order interaction. For example, the saturated four-factor model would be specified by entering "*ABCD*."

Step-Down Search Options

Perform Step-Down Search

Specifies whether the step-down model selection procedure is used. The procedure starts with the model specified in the Model option. The procedure is controlled by the Max Models and the Stopping Alpha options as explained below.

Note that the step-down search is a lengthy procedure and may require a few minutes to execute.

Max Models

This option specifies the maximum number of models that can be tried before the search is aborted. On slower computers, the search may go on for hours, so this option lets you abort the search after so many iterations. On Pentium class computers, the search will take only a few moments, so you can set this option very high.

Stopping Alpha

This option specifies the value of alpha which is the significance level for the goodness of fit tests. During the search, when no model is found whose probability level is greater than this amount, the search is ended. Remember that you are searching for a model the fits the data well and thus does not produce a significant goodness of fit test.

Although you might be in the habit of always selecting an alpha level of 0.05, you should consider using a larger value (say 0.15 or 0.25) because you want a model that fits the data well--that is not even close to significance. A model that is almost significant (has an alpha of 0.06 or 0.08) might be excluding important terms. When you use a value of 0.25, you can feel confident that your model really fits the data well.

Unfortunately, the appropriate value of alpha is also tied to the sample size. With small samples, a significance level of 0.25 might be due to a lack of the goodness of fit test's ability to reject any hypothesis and not a general agreement between the model and the data. Hence, with small sample sizes you can have a poor fit and a high alpha. On the other hand, large sample sizes may cause even slight deviations between model and fit to be significant at the 0.05 level. Hence, for large sample sizes, you would want the value of alpha to be closer to zero.

Maximum Likelihood Options

Max Iterations

This option specifies the maximum number of iterations. Usually, the algorithm will converge in less than five iterations, so the default value of twenty-five should be more than ample.

Max Difference

This option specifies the maximum difference between any of the actual frequencies and their corresponding predicted frequencies. Once the maximum is less than this amount, the maximum-likelihood procedure will terminate (has converged).

Reports Tab

The following options control which reports are displayed. Note that some reports are not produced in certain situations. If a report you want is not created, try rerunning the program with a specific model--with all Model Selection reports turned off.

Select Reports

Multi-Term Report - Table Report

Specifies whether the report is produced.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Value Labels

Indicate whether to display the data values or their labels.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Loglinear Model Analysis

This section presents an example of how to run an analysis of the data contained in the LOGLIN1 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Loglinear Models window.

1 Open the LOGLIN1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Loglin1.s0**.
- Click **Open**.

2 Open the Loglinear Models window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Loglinear Models**. The Loglinear Models procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Loglinear Models window, select the **Variables tab**.
- Double-click in the **Factor A Variable** text box. This will bring up the variable selection window.
- Select **Newspaper** from the list of variables and then click **Ok**. "Newspaper" will appear in the Factor A Variables box.
- Double-click in the **Factor B Variable** text box. This will bring up the variable selection window.
- Select **Lecture** from the list of variables and then click **Ok**. "Lecture" will appear in the Factor B Variables box.
- Double-click in the **Factor C Variable** text box. This will bring up the variable selection window.
- Select **Radio** from the list of variables and then click **Ok**. "Radio" will appear in the Factor C Variables box.
- Double-click in the **Factor D Variable** text box. This will bring up the variable selection window.
- Select **Reading** from the list of variables and then click **Ok**. "Reading" will appear in the Factor D Variables box.
- Double-click in the **Factor E Variable** text box. This will bring up the variable selection window.
- Select **Knowledge** from the list of variables and then click **Ok**. "Knowledge" will appear in the Factor E Variables box.
- Double-click in the **Frequency Variable** text box. This will bring up the variable selection window.
- Select **Counts** from the list of variables and then click **Ok**. "Counts" will appear in the Frequency Variable box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Multiple-Term Test Section

Multiple-Term Test Section

K-Terms	DF	Like. Ratio Chi-Square	Prob Level	Pearson Chi-Square	Prob Level
1WAY & Higher	31	2666.19	0.0000	3811.81	0.0000
2WAY & Higher	26	596.84	0.0000	751.31	0.0000
3WAY & Higher	16	19.56	0.2406	21.21	0.1705
4WAY & Higher	6	3.23	0.7791	3.32	0.7680
5WAY & Higher	1	1.02	0.3116	1.01	0.3157

Note: Fit of all k-factor marginals. Simultaneous test that all interactions of order k and higher are zero.

K-Terms	DF	Like. Ratio Chi-Square	Prob Level
1WAY Only	5	2069.35	0.0000
2WAY Only	10	577.28	0.0000
3WAY Only	10	16.33	0.0906
4WAY Only	5	2.21	0.8196

Note: Simultaneous test that all interactions of order k are zero. These Chi-Squares are differences in the above table.

This report helps in the model selection process by isolating the highest order term(s) that need to be included in the final LLM.

The top table shows the significance of all terms of a given order and higher. For example, the 596.84 tests the significance of all terms of order two and above. The 3.23 tests the significance of all fourth- and fifth-order terms. Since there are only five factors in the table, the 1.02 tests the significance of the five-way interaction.

By glancing down the significance levels (Prob Level) of this table, you can quickly determine the maximum order that is significant. In the present example, note that the one-way and higher is significant, as is the two-way and higher. However, the three-way and higher is not significant, being only 0.2406 (we use a significance level of 0.20). Hence, all terms of order three or greater may be ignored.

The second table is formed by differencing the first. Since the Pearson chi-square cannot be differenced in this manner, only the likelihood-ratio chi-square tests are shown. These tests indicate the significance of all terms of a given order. They are used to substantiate the conclusions made from the first table.

In this example, you notice that the four-way and three-way terms are not significant, while the two-way and one-way terms are. Again, we are lead to the conclusion that second-order terms will be the highest that are needed in our final model.

Individual definitions of the columns of this report are as follows:

K-Terms

These are the terms that are being tested. In the first table they are the terms that are not in the model. Hence the goodness-of-fit chi-square test indicates whether these terms may be left out of the model. In the second table, these are the terms being tested.

DF

The degrees of freedom of the terms being tested. This is a parameter of the chi-square distribution. The degrees of freedom of the test are found by adding up the degrees of freedom of the individual terms left out of the model.

Like. Ratio Chi-Square

This is the value of the likelihood-ratio statistic calculated using the following formula:

$$G^2 = 2\sum_{i,j,k} f_{ijk} \ln\left(\frac{f_{ijk}}{\hat{m}_{ijk}}\right)$$

This statistic follows the chi-square distribution in moderate to large samples. It is calculated using this formula in the top half of the report and by subtracting one row from the previous row in the bottom half of the report.

Note that strictly speaking, the likelihood-ratio statistics in the second table follow the central chi-square distribution only if the second chi-square (the one subtracted) is not significant.

Prob Level

This is the probability of obtaining the above chi-square value or larger by chance. When this value is less than some preset alpha level, say 0.15, the test statistic is said to be *significant*. Otherwise, the test statistic is *nonsignificant*. A nonsignificant model fits the data adequately. The choice of 0.15 is arbitrary, and you may use whatever value you feel comfortable with between 0.300 and 0.001.

Pearson Chi-Square

This is the value of the Pearson chi-square statistic calculated using the following formula,

$$\chi^2 = 2\sum_{i,j,k} \frac{\left(f_{ijk} - \hat{m}_{ijk}\right)^2}{\hat{m}_{ijk}}$$

Prob Level

This is the probability of obtaining the above chi-square value or larger by chance. When this value is less than some preset alpha level, say 0.15, the test statistic is said to be *significant*. Otherwise, the test statistic is *nonsignificant*. A nonsignificant model fits the data adequately. The choice of 0.15 is arbitrary, and you may use whatever value you feel comfortable with between 0.300 and 0.001.

Single-Term Test Section

Single-Term Test Section

Effect	DF	Partial Chi-Square	Prob Level	Marginal Chi-Square	Prob Level
A (Newspaper)	1	27.31	0.0000	27.31	0.0000
B (Lecture)	1	1449.22	0.0000	1449.22	0.0000
C (Radio)	1	498.12	0.0000	498.12	0.0000
D (Reading)	1	4.58	0.0323	4.58	0.0323
E (Knowledge)	1	90.11	0.0000	90.11	0.0000
AB	1	4.52	0.0335	21.60	0.0000
AC	1	46.15	0.0000	74.23	0.0000
AD	1	172.49	0.0000	253.71	0.0000
AE	1	31.75	0.0000	105.78	0.0000
BC	1	10.27	0.0014	18.95	0.0000
BD	1	7.33	0.0068	23.75	0.0000
BE	1	4.82	0.0282	17.16	0.0000
CD	1	0.42	0.5147	21.08	0.0000
CE	1	6.28	0.0122	24.25	0.0000
DE	1	79.58	0.0000	150.45	0.0000
ABC	1	1.48	0.2240	1.85	0.1733
ABD	1	0.01	0.9418	0.36	0.5470
ABE	1	3.16	0.0755	1.16	0.2820
ACD	1	1.25	0.2642	1.81	0.1790
ACE	1	0.00	0.9526	0.06	0.8089
ADE	1	2.75	0.0971	3.06	0.0800
BCD	1	1.50	0.2208	3.48	0.0621
BCE	1	1.39	0.2377	0.51	0.4740
BDE	1	3.86	0.0494	4.30	0.0381
CDE	1	0.01	0.9435	0.57	0.4486

This report presents partial and marginal association tests on terms of up to the third order. The actual computation was discussed earlier in the section of model selection, so we will not repeat it here except to note that the chi-squares are the difference between the likelihood-ratio statistics of two models. The validity of this procedure depends on the more complex model's likelihood ratio being nonsignificant.

You should remember that the partial chi-square statistic tests whether the term is significant after considering all other terms of the same order. The marginal-association chi-square tests whether the term is significant ignoring all other terms of the same order. Hence, when both tests are significant, you can be fairly certain that the term is necessary. When neither test is significant, you can be fairly certain that the term is not necessary. And when one test is significant and the other is not, the term should be watched closely--it may or may not be important.

In this example, notice that only one three-way term, *BDE*, is significant. Almost all of the two-way terms are significant. Hence, our search for the best model might begin with the hierarchical model: *AB,AC,AD,AE,BC,BDE,CD,CE*. The *CD* term was not significant on the partial association test, so we might expect to see it omitted from the final model.

Notice that even though the simultaneous test of all third-order terms was not significant, this report indicated that *BDE* should be considered. There is always a possibility of this type of confusion among the various goodness of fit tests. This is why it is important to look at all of them. You can rationalize the difference in conclusions here by noting that the *BDE* is not highly significant, but only mildly significant.

Step-Down Model Search Section

Step-Down Model-Search Section

Step No	Best No	DF	Chi-Square	Prob Level	Term Deleted	DF	Chi-Square	Prob Level	Hierarchical Model
1	1	0	0.0	1.0000	None	0	0.0	0.0000	ABCDE
2	1	1	1.0	0.3116	ABCDE	1	1.0	0.3116	BCDE,ACDE,ABDE,ABCE,ABCD
3	2	2	1.3	0.5096	BCDE	1	0.3	0.5690	ACDE,ABDE,ABCE,ABCD
4	2	2	1.2	0.5463	ACDE	1	0.2	0.6670	BCDE,ABDE,ABCE,ABCD
5	2	2	1.8	0.4074	ABDE	1	0.8	0.3796	BCDE,ACDE,ABCE,ABCD
6	2	2	1.2	0.5598	ABCE	1	0.1	0.7117	BCDE,ACDE,ABDE,ABCD
7	2	2	1.4	0.4935	ABCD	1	0.4	0.5330	BCDE,ACDE,ABDE,ABCE
8	6	3	1.7	0.6464	BCDE	1	0.5	0.4808	ACDE,ABDE,ABCD,BCE
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
98	86	18	24.4	0.1420	AB	1	4.4	0.0354	CE,AC,BC,AD,AE,DE,BE,BD
99	86	18	100.7	0.0000	DE	1	80.7	0.0000	CE,AC,BC,AD,AE,AB,BE,BD
100	86	18	24.7	0.1324	BE	1	4.7	0.0293	CE,AC,BC,AD,AE,AB,DE,BD
101	86	18	27.7	0.0674	BD	1	7.7	0.0056	CE,AC,BC,AD,AE,AB,DE,BE
102	98	19	31.5	0.0360	CE	1	7.0	0.0079	AC,BC,AD,AE,DE,BE,BD
103	98	19	81.6	0.0000	AC	1	57.2	0.0000	CE,BC,AD,AE,DE,BE,BD
104	98	19	37.8	0.0064	BC	1	13.4	0.0003	CE,AC,AD,AE,DE,BE,BD
105	98	19	210.5	0.0000	AD	1	186.1	0.0000	CE,AC,BC,AE,DE,BE,BD
106	98	19	57.3	0.0000	AE	1	32.9	0.0000	CE,AC,BC,AD,DE,BE,BD
107	98	19	103.8	0.0000	DE	1	79.4	0.0000	CE,AC,BC,AD,AE,BE,BD
108	98	19	30.6	0.0449	BE	1	6.2	0.0130	CE,AC,BC,AD,AE,DE,BD
109	98	19	37.4	0.0070	BD	1	13.0	0.0003	CE,AC,BC,AD,AE,DE,BE
Best model found: CE,AC,BC,AD,AE,DE,BE,BD									
98	98	18	24.4	0.1420	AB	1	4.4	0.0354	CE,AC,BC,AD,AE,DE,BE,BD

This report documents the search algorithm's progress. It shows the results of each step. Remember that the algorithm uses a step down strategy. This means that it begins with the most complicated model possible (the saturated model) and proceeds by removing terms. Your main interest will be in the final model selected, but sometimes it is of interest to see how this model was arrived at.

Step No

This is the identification number of this model. This is the number referred to under Best.

Best No

The number of the model that is currently the designated as being the best.

DF

The degrees of freedom of the chi-square value. This is the degrees of freedom of the terms not in the model, since these are the terms being tested.

Chi-Square

The likelihood-ratio statistic, G^2, testing the goodness of fit of this model. This statistic tests the significance of the terms omitted from the model. Hence, when the G^2 is not significant, you can assume that all important terms are in the model. Of course, you might have included some negligible terms as well.

Prob Level

This is the probability value for the above chi-square statistic. If it is less than some small value, say 0.05, the chi-square is said to be significant and you assume that one of the terms left out of

the model is important. If the probability is greater than the cutoff value, you assume that all significant terms are accounted for.

Term Deleted

This is the term that was removed from the current “best” model to obtain this model. Note that the model is reduced by that term only and not by all terms of lower order that were included because of it.

DF

The degrees of freedom of the term removed.

Chi Square

This value tests the significance of the removed term. It is calculated as the difference between the current chi-square statistic and the current best model’s chi-square statistic. Since these are nested likelihood-ratio statistics, this difference is also a chi-square statistic.

Prob Level

The probability of rejecting the above chi-square value. If this value is greater than 0.05, you can assume that term is not necessary in the model.

Hierarchical Model

This is the hierarchical model that was fit.

Model Section

Model Section

Hierarchical Model: CE,AC,BC,AD,AE,DE,BE,BD

Model Term	Individual DF	Cumulative DF
Mean	1	1
A	1	2
B	1	3
C	1	4
AC	1	5
BC	1	6
D	1	7
AD	1	8
BD	1	9
E	1	10
AE	1	11
BE	1	12
CE	1	13
DE	1	14
Error	18	32

This report presents the expanded model (all terms are listed) as well as the associated degrees of freedom.

Chi-Square Test Section

Chi-Square Tests Section

DF	Like. Ratio Chi-Square	Prob Level	Pearson Chi-Square	Prob Level	Model
18	24.41	0.1420	24.49	0.1395	CE,AC,BC,AD,AE,DE,BE,BD

This report presents details of both the likelihood-ratio and the Pearson chi-square goodness of fit tests of the model selected. These terms are defined above.

Parameter Estimation Section

Parameter Estimation Section

Model Term	Number Cells	Count	Percent Count	Average Log(Count)	Effect (Lambda)	Effect Std. Error	Effect Z-Value
Mean	32	1729	100.00	3.0186	3.0186	0.0598	50.48
A: Newspaper							
1	16	973	56.28	3.3620	0.3434	0.0598	5.74
2	16	756	43.72	2.6752	-0.3434	0.0598	-5.74
B: Lecture							
1	16	135	7.81	1.8370	-1.1816	0.0598	-19.76
2	16	1594	92.19	4.2001	1.1816	0.0598	19.76
C: Radio							
1	16	412	23.83	2.5328	-0.4858	0.0598	-8.12
2	16	1317	76.17	3.5044	0.4858	0.0598	8.12
D: Reading							
1	16	820	47.43	3.1129	0.0944	0.0598	1.58
2	16	909	52.57	2.9242	-0.0944	0.0598	-1.58
E: Knowledge							
1	16	668	38.64	2.8902	-0.1283	0.0598	-2.15
2	16	1061	61.36	3.1469	0.1283	0.0598	2.15
AC: Newspaper,Radio							
1,1	8	306	17.70	3.1110	0.2349	0.0598	3.93
1,2	8	667	38.58	3.6129	-0.2349	0.0598	-3.93
2,1	8	106	6.13	1.9545	-0.2349	0.0598	-3.93
2,2	8	650	37.59	3.3959	0.2349	0.0598	3.93
BC: Lecture,Radio							
1,1	8	54	3.12	1.5274	0.1762	0.0598	2.95
1,2	8	81	4.68	2.1467	-0.1762	0.0598	-2.95
2,1	8	358	20.71	3.5381	-0.1762	0.0598	-2.95
2,2	8	1236	71.49	4.8622	0.1762	0.0598	2.95

(Report continues)

This report provides the details of the loglinear estimation of the specified model. This report was the goal of the LLM analysis. The definitions are as follows:

Model Term

The particular term in the model. Note that the levels of the term are also listed below the term. These levels would have printed out in words (like YES and NO) if Value Labels option had been set appropriately.

Number Cells

The number of cells involved in this term.

Count

The total of all cell counts at the indicated levels.

Percent Count

The percent the Count is of the table total. These percentages are used to understand why the term was significant.

Average Log(Count)

The average of LOG(count+delta) of all cells at the indicated levels.

Effect (Lambda)

The estimated value of λ for this term. These $\lambda's$ were identified in above and estimated using the routine of Haberman (1972).

Effect Std. Error

The asymptotic standard error of the above effect. When a saturated model is fit, the standard error is given by the square root of the variance of the effect. The variance is estimated using the formulas provided in Lee (1977). When an incomplete (less than saturated) model is estimated, the program uses the resulting estimated cell counts in the formulas appropriate for the saturated models. This is called the *approximate method* by Lee (1977). He states that these estimates may be a little large.

Effect Z-Value

This is the effect divided by the standard error. Since the number of cells included in a term differs from term to term, their estimation precision also differs. This z-value allows you to compare the relative magnitudes of the effects across all main-effects and interactions. These values represent the relative importance of that term in the loglinear model. The term *z-value* is used because these values are asymptotically normal. These were called the *standardized parameter estimates* in the Model Selection section of this chapter (presented earlier).

One model-selection rule of thumb is that you should keep all terms which have at least one z-value greater than some cutoff value (say 2.0 or 3.0) in absolute value.

Interpreting Significant Effects

The final task in loglinear analysis involves interpreting a significant term. This is usually accomplished by collapsing the table to the factors in the term of interest and then analyzing the percentages. For example, the term BC was significant. From above report, we can construct the following two-way table of percentages from the Percent Count column of the report. Note that we have arbitrarily decided to sum across the table. You could have summed down the table instead with no loss in analysis capability.

	Lectures		
Radio	**Yes (B=1)**	**No (B=2)**	**Total**
Yes (C=1)	**13%** = 100(3.12/23.83)	**87%** = 100(20.71/23.83)	**100%** = 100(23.83/23.83)
No (C=2)	**6%** = 100(4.68/76.17)	**94%** = 100(71.49/76.17)	**100%** = 100(76.17/76.17)

Looking at these percentages, we can now see why this term was significant. Notice that when factor C is 1, factor B changes from 13% to 87%. However, when factor C is 2, factor B changes from 6% to 94%. This difference in the amount of change is what causes BC to be significant. This type of table should be created for every significant term.

Data Table Section

Data Table Section

Reading	Radio	Lecture	Newspaper	Actual	Pred	Diff	Chi	FT-SR
Knowledge=1								
1	1	1	1	23.0	24.9	-1.9	-0.39	-0.34
1	1	1	2	1.0	2.7	-1.7	-1.05	-1.04
1	1	2	1	102.0	103.6	-1.6	-0.15	-0.13
1	1	2	2	16.0	11.4	4.6	1.38	1.31
1	2	1	1	27.0	24.7	2.3	0.47	0.50
1	2	1	2	3.0	6.9	-3.9	-1.49	-1.62
1	2	2	1	201.0	207.6	-6.6	-0.46	-0.44
1	2	2	2	67.0	58.2	8.8	1.15	1.14
2	1	1	1	8.0	4.3	3.7	1.77	1.55
2	1	1	2	4.0	2.0	2.0	1.39	1.22
2	1	2	1	35.0	36.2	-1.2	-0.19	-0.15
2	1	2	2	13.0	16.9	-3.9	-0.95	-0.94
2	2	1	1	7.0	4.3	2.7	1.32	1.22
2	2	1	2	2.0	5.1	-3.1	-1.38	-1.49
2	2	2	1	75.0	72.5	2.5	0.30	0.32
2	2	2	2	84.0	86.7	-2.7	-0.29	-0.27
Knowledge=2								
1	1	1	1	8.0	9.6	-1.6	-0.51	-0.44
1	1	1	2	3.0	2.0	1.0	0.68	0.71
1	1	2	1	67.0	63.8	3.2	0.40	0.42
1	1	2	2	16.0	13.5	2.5	0.68	0.71
1	2	1	1	18.0	13.0	5.0	1.37	1.31
1	2	1	2	8.0	7.1	0.9	0.35	0.42
1	2	2	1	177.0	175.8	1.2	0.09	0.11
1	2	2	2	83.0	95.2	-12.2	-1.25	-1.26
2	1	1	1	4.0	4.4	-0.4	-0.20	-0.08
2	1	1	2	3.0	4.0	-1.0	-0.49	-0.38
2	1	2	1	59.0	59.2	-0.2	-0.03	0.00
2	1	2	2	50.0	53.5	-3.5	-0.47	-0.45
2	2	1	1	6.0	6.0	0.0	0.00	0.09
2	2	1	2	10.0	13.9	-3.9	-1.04	-1.04
2	2	2	1	156.0	163.1	-7.1	-0.56	-0.55
2	2	2	2	393.0	376.9	16.1	0.83	0.83

This report presents the cell counts along with their predicted values and residuals. The main purpose of this report is to let you look for large residuals--cells that are predicted poorly by the LLM.

Actual

The cell count f_{ijk} which was read in or tabulated from the database.

Predicted

The predicted cell count m_{ijk} based on the current hierarchical model. The prediction equation is of the following form, with estimation by maximum likelihood.

$$\ln\left(m_{ijk}\right) = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

The algorithm of Haberman (1972) is used to produce the maximum-likelihood estimates.

Difference

These residuals are Actual - Predicted. They are usually scanned to find cells that are not fit well by the model. Since the size of a residual must be judged in terms of the relative size of the cell count, you should avoid simply finding the largest residuals. Instead, you should look at a standardized residual, such as the Chi value in the last column.

Chi

This is a standardized residual. It is calculated using the formula

$$Chi = \frac{f_{ijk} - m_{ijk}}{\sqrt{m_{ijk}}}$$

It is the square root of the contribution of this cell to the overall Pearson chi-square goodness of fit statistic. This standardized residual lets you make direct comparisons among the fits of the various cells. Values of Chi larger than 1.96 in absolute value would be considered large.

FT-SR

This is the Freeman-Tukey standardized residual. Freeman and Tukey pointed out that for observations from a Poisson distribution, the quantity $\sqrt{x} + \sqrt{x+1}$ has a mean approximately equal to $\sqrt{4\mu+1}$ and a variance of one. Using this result, they formed this statistic which is written in our notation as

$$FTSR = \sqrt{f_{ijk}} + \sqrt{f_{ijk}+1} - \sqrt{4m_{ijk}+1}$$

Notice that this value does not suffer when the denominator is zero which is a real difficulty with the Chi statistic.

This value may also be considered as being from the unit normal distribution. Hence, like Chi, absolute values greater than 1.96 are considered larger.

Chapter 535

Binary Diagnostic Tests – Single Sample

Introduction

An important task in diagnostic medicine is to measure the accuracy of a diagnostic test. This can be done by comparing the test result with the true condition status of a number of patients. The results of such a study can be displayed in a 2-by-2 table in which the true condition is shown as the rows and the diagnostic test result is shown as the columns.

	Diagnostic Test Result		
True Condition	Positive	Negative	Total
Present (True)	*T1*	*T0*	*n1*
Absent (False)	*F1*	*F0*	*n0*
Total	*m1*	*m0*	*N*

Data such as this can be analyzed using the standard techniques for two proportions. However, specialized techniques have been developed for dealing specifically with the questions that arise from such a study. These techniques are presented in the book by Zhou, Obuchowski, and McClish (2002), and this is the reference that we have used in developing this procedure.

Test Accuracy

Several measures of a diagnostic test's accuracy are available. Probably the most popular measures are the test's *sensitivity* and the *specificity*. Sensitivity is the proportion of those that have the condition for which the diagnostic test is positive. Specificity is the proportion of those that do not have the condition for which the diagnostic test is negative. Other accuracy measures that have been proposed are the likelihood ratio and the odds ratio.

Technical Details

Suppose you arrange the results of a diagnostic test into a 2-by-2 table as follows:

	Diagnostic Test Result		
True Condition	Positive	Negative	Total
Present (True)	*T1*	*T0*	*n1*
Absent (False)	*F1*	*F0*	*n0*
Total	*m1*	*m0*	*N*

Sensitivity and Specificity

The sensitivity is estimated as

$$\hat{Se} = \frac{T1}{n1}$$

and the specificity is estimated as

$$\hat{Sp} = \frac{F0}{n0}$$

Confidence intervals may be formed for these two statistics. Rather than use the common confidence interval for a proportion that uses the normal approximation to the binomial, we use the more accurate score method of Wilson (1927). This method has been shown by Agresti and Coull (1998) to have much better coverage probabilities than either the exact method of inverting the binomial or the simple Wald confidence interval.

The confidence limits for the sensitivity based on the score method are

$$\frac{\hat{Se} + \frac{z^2_{1-\alpha/2}}{2n1} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{Se}\left(1-\hat{Se}\right) + \frac{z^2_{1-\alpha/2}}{4n1}}{n1}}}{1 + \frac{z^2_{1-\alpha/2}}{n1}}$$

and for specificity are

$$\frac{\hat{Sp} + \frac{z^2_{1-\alpha/2}}{2n0} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{Sp}\left(1-\hat{Sp}\right) + \frac{z^2_{1-\alpha/2}}{4n0}}{n0}}}{1 + \frac{z^2_{1-\alpha/2}}{n0}}$$

Likelihood Ratio

The likelihood ratio (LR) statistic may be used as a measure of accuracy of a diagnostic test. This statistic is calculated both for positive and negative test results as follows

$$\begin{aligned} LR(+) &= \frac{P(\text{Test} = \text{Positive}|\text{Condition} = \text{Present})}{P(\text{Test} = \text{Positive}|\text{Condition} = \text{Absent})} \\ &= \frac{Se}{1-Sp} \\ &= \frac{Se}{FPR} \end{aligned}$$

and

$$\begin{aligned} LR(-) &= \frac{P(\text{Test} = \text{Negative}|\text{Condition} = \text{Present})}{P(\text{Test} = \text{Negative}|\text{Condition} = \text{Absent})} \\ &= \frac{1-Se}{Sp} \\ &= \frac{FPR}{Sp} \end{aligned}$$

where *FPR* is the false positive rate and *FNR* is the false negative rate.

Confidence limits for LR(+)are calculated using the skewness adjusted score method of Gart and Nam (1998). The lower limit is the solution of

$$\left(\frac{S(\phi, \tilde{p}_2)}{\sqrt{V}} - \frac{\mu_3(z^2-1)}{6} \right)^2 - z^2 = 0$$

and the upper limit is the solution of

$$\left(\frac{S(\phi, \tilde{p}_2)}{\sqrt{V}} + \frac{\mu_3(z^2-1)}{6} \right)^2 - z^2 = 0$$

where $\tilde{p}_2$ is the appropriate solution of

$$N\tilde{p}_2^2 - \left[\phi(n1+T1) + F1 + n1\right]\tilde{p}_2 + m1 = 0$$

and

$$S(\phi, \tilde{p}_2) = \frac{T1 - n1(\tilde{p}_1)}{\phi\tilde{q}_1}$$

$$V = \left(\phi^2 \left(\frac{\tilde{q}_1}{(n1)\tilde{p}_1} + \frac{\tilde{q}_2}{n0(\tilde{p}_2)} \right) \right)^{-1}$$

$$\tilde{p}_1 = \phi\tilde{p}_2$$

$$\tilde{q}_1 = 1 - \tilde{p}_1$$

$$\tilde{q}_2 = 1 - \tilde{p}_2$$

$$\tilde{\mu}_3 = v^{3/2}\left(\frac{\tilde{q}_1(\tilde{q}_1 - \tilde{p}_1)}{\left(n1(\tilde{p}_1)\right)^2} - \frac{\tilde{q}_2(\tilde{q}_2 - \tilde{p}_2)}{\left(n0(\tilde{p}_2)\right)^2}\right)$$

$$v = \left(\frac{\tilde{q}_1}{n1(\tilde{p}_1)} + \frac{\tilde{q}_2}{n0(\tilde{p}_2)}\right)^{-1}$$

Using the substitution

$$\begin{aligned}\phi &= LR(+) \\ &= \frac{Se}{1 - Sp} \\ &= \frac{p_1}{p_2}\end{aligned}$$

The formulas for LR(-) are similar. They are based on the substitution

$$\begin{aligned}\phi &= LR(-) \\ &= \frac{1 - Se}{Sp} \\ &= \frac{p_1}{p_2}\end{aligned}$$

Odds Ratio

Another measure of accuracy is the odds ratio which is

$$o = \frac{\left(\frac{Se}{1 - Se}\right)}{\left(\frac{1 - Sp}{Sp}\right)}$$

Formulas for computing confidence limits of the odds ratio are given in the chapter on Two Proportions and they will not be repeated here.

Data Structure

This procedure does not use data from the database. Instead, you enter the values directly into the panel. The data are entered in the familiar 2-by-2 table format.

Procedure Options

This section describes the options available in this procedure.

Data Tab

Enter the data values directly on this panel.

Data Values

T1

This is the number of patients that had the condition of interest and responded positively to the diagnostic test.

T0

This is the number of patients that had the condition of interest but responded negatively to the diagnostic test.

F1

This is the number of patients that did not have the condition of interest but responded positively to the diagnostic test.

F0

This is the number of patients that did not have the condition of interest and responded negatively to the diagnostic test.

Report Options

Alpha - Confidence Limits

The confidence coefficient to use for the confidence limits of the difference in proportions. 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5.

Decimal - Proportions

The number of digits to the right of the decimal place to display when showing proportions on the reports.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Binary Diagnostic Test of a Single Sample

This section presents an example of how to run an analysis on hypothetical data. In this example, samples of 50 individuals known to have a certain disease and 50 individuals without the disease where selected at random. All 100 individuals were given a diagnostic test. Of those with the disease, 42 tested positively and 8 tested negatively for it on the diagnostic test. Of those without the disease, 14 tested positively and 36 tested negatively for it on the diagnostic test.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Binary Diagnostic Tests – Single Sample window.

1 Open the Binary Diagnostic Tests – Single Sample window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **Binary - Single Sample**. The Binary Diagnostic Tests – Single Sample procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the data.

- On the Binary Diagnostic Tests – Single Sample window, select the **Data tab**.
- In the **T1** box, enter **42**.
- In the **T0** box, enter **8**.
- In the **F1** box, enter **14**.
- In the **F0** box, enter **36**.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Table and Data Sections

Counts

True Condition	Diagnostic Test Result Positive	Negative	Total
Present	42	8	50
Absent	14	36	50
Total	56	44	100

Table Proportions

	Diagnostic Test Result Positive	Negative	Total
	0.4200	0.0800	0.5000
	0.1400	0.3600	0.5000
	0.5600	0.4400	1.0000

Row Proportions

True Condition	Diagnostic Test Result Positive	Negative	Total
Present	0.8400	0.1600	1.0000
Absent	0.2800	0.7200	1.0000
Total	0.5600	0.4400	1.0000

Column Proportions

	Diagnostic Test Result Positive	Negative	Total
	0.7500	0.1818	0.5000
	0.2500	0.8182	0.5000
	1.0000	1.0000	1.0000

These reports display the data table that was input along with various proportions that make interpreting the table easier. Note that the sensitivity and specificity are displayed in the Row Proportions table.

Sensitivity and Specificity Section

Measure	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit	Number Matches	Number Different	Total
Sensitivity	0.8400	0.7149	0.9166	42	8	50
Specificity	0.7200	0.5833	0.8253	36	14	50

This report displays the sensitivity and specificity with their corresponding confidence limits. Note that for a perfect diagnostic test, both values would be one. Hence, the higher the values the better.

Likelihood Ratio Section

Measure	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
LR(Test=Posivitive)	3.0000	1.9753	5.0086
LR(Test=Negative)	0.2222	0.1062	0.4026

This report displays LR(+) and LR(-) with their corresponding confidence limits. You would want LR(+) > 1 and LR(-) < 1, so you should place close attention that the lower limit of LR(+) is greater than one and that the upper limit of LR(-) is less than one.

Note the LR(+) means LR(Test=Positive). Similarly, LR(-) means LR(Test=Negative).

Odds Ratio Section

Measure	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Odds Ratio (+ 1/2)	12.5862	4.8421	32.7155
Odds Ratio (Fleiss)	12.5862	4.6339	40.8823

This report displays estimates of the odds ratio as well as confidence limits for the odds ratio. Because of the better coverage probabilities of the Fleiss confidence interval, we suggest that you use the second line of the report.

Chapter 536

Binary Diagnostic Tests – Paired Samples

Introduction

An important task in diagnostic medicine is to measure the accuracy of two diagnostic tests. This can be done by comparing summary measures of diagnostic accuracy such as *sensitivity* or *specificity* using a statistical test. Often, you want to show that a new test is similar to another test, in which case you use an equivalence test. Or, you may wish to show that a new diagnostic test is not inferior to the existing test, so you use a noninferiority test. All of these hypothesis tests are available in this procedure for the important case when the diagnostic tests provide a binary (yes or no) result.

Experimental Design

Suppose you are interested in comparing the sensitivities of two diagnostic tests for a particular disease (or condition). Each test provides a binary (yes or no) response. Further suppose you draw a random sample of subjects from the population with the disease and administered both diagnostic tests to each subject in random order. Assume that Test 1 is a new (experimental or treatment) test that will replace Test 2, the existing (standard or reference) test, if it is found to be better.

The results of such a study can be displayed in a 2-by-2 table in which the Test 1 result is shown as the rows and the Test 2 result is shown as the columns.

	Test 2 Result		
Test 1 Result	Positive	Negative	Total
Positive	*X11*	*X10*	*m1*
Negative	*X01*	*X00*	*m0*
Total	*n1*	*n0*	*N*

Data such as this can be analyzed using standard techniques for comparing two correlated proportions which are presented in the chapter on Two Correlated Proportions. Such a table was

originally analyzed using McNemar's Test. However, procedures with better statistical properties have recently been proposed. See for example Nam and Blackwelder (2002).

Sensitivity

Sensitivity is the proportion of those that have the condition for which the diagnostic test is positive. Since this design assumes that the subjects come from the population of individuals with the disease, the sensitivity can be calculated.

Specificity

Specificity is the proportion of those that do not have the condition for which the diagnostic test is negative. To study specificity, a separate study would have to be conducted in which subjects were drawn from the population of individuals without the disease. The data from a such a study could be analyzed with this procedure by changing the meaning of *positive* and *negative*. Instead of positive meaning that the person had the disease, positive would mean that the diagnostic test result matched the true condition of the subject. Likewise, negative would mean that the diagnostic test result did not match the true condition. In the procedure printouts, you would substitute specificity for sensitivity.

Comparing Sensitivity and Specificity

Suppose you arrange the results of two diagnostic tests into two 2-by-2 tables as follows:

	Test 2 Result		
Test 1 Result	Positive	Negative	Total
Positive	*X11*	*X10*	*m1*
Negative	*X01*	*X00*	*m0*
Total	*n1*	*n0*	*N*

Hence, the study design include $N = N1 + N0$ patients.

The hypotheses of interest when comparing the sensitivities (Se) of two diagnostic tests are either the difference hypotheses

$$H_O: Se1 - Se2 = 0 \text{ versus } H_A: Se1 - Se2 \neq 0$$

or the ratio hypothesis

$$H_O: Se1 / Se2 = 1 \text{ versus } H_A: Se1 / Se2 \neq 1$$

Similar sets of hypotheses may be defined for the difference or ratio of the specificities (Sp) as

$$H_O: Sp1 - Sp2 = 0 \text{ versus } H_A: Sp1 - Sp2 \neq 0$$

and

$$H_O: Sp1 / Sp2 = 1 \text{ versus } H_A: Sp1 / Sp2 \neq 1$$

Note that the difference hypotheses usually require a smaller sample size for comparable statistical power, but the ratio hypotheses may be more convenient.

The sensitivities are estimated as

$$\hat{Se}1 = \frac{X11}{m1} \text{ and } \hat{Se}2 = \frac{X01}{n1}$$

The sensitivities of the two diagnostic tests may be compared using either their differences or their ratios. Hence, the comparison of the sensitivity reduces to the problem of comparing two correlated binomial proportions. The formulas used for hypothesis testing and confidence intervals are the same as presented in the chapter on testing two correlated proportions. We refer you to that chapter for further details.

Data Structure

This procedure does not use data from the database. Instead, you enter the values directly into the 2-by-2 table on the panel.

Procedure Options

This section describes the options available in this procedure.

Data Tab

Enter the data values directly on this panel.

Data Values

X11

This is the number of patients that responded positively to both diagnostic tests. The value entered must be a non-negative number.

X10

This is the number of patients that tested positive using Test 1, but negative using Test 2. The value entered must be a non-negative number.

X01

This is the number of patients that tested negative using Test 1, but positive using Test 2. The value entered must be a non-negative number.

X00

This is the number of patients that responded negatively to both diagnostic tests. The value entered must be a non-negative number.

Confidence Interval Method

Difference C.I. Method

This option specifies the method used to calculate the confidence intervals of the sensitivity differences. These methods are documented in detail in the Two Correlated Proportions chapter. We recommend the score method proposed by Nam (1990).

Ratio C.I. Method

This option specifies the method used to calculate the confidence intervals of the sensitivity ratios. These methods are documented in detail in the Two Correlated Proportions chapter. The recommended method is score method proposed by Nam and Blackwelder (2002).

Report Options

Alpha - Confidence Intervals

The confidence coefficient to use for calculating the confidence limits in proportions. 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05.

Alpha - Hypothesis Tests

This is the significance level of the hypothesis tests, including the equivalence and noninferiority tests. Typical values are between 0.01 and 0.10. The most common choice is 0.05.

Proportion Decimals

The number of digits to the right of the decimal place to display when showing proportions on the reports.

Probability Decimals

The number of digits to the right of the decimal place to display when showing probabilities on the reports.

Equivalence or Non-Inferiority Settings

Max Equivalence Difference

This is the largest value of the difference between the two sensitivities that will still result in the conclusion of equivalence. When running equivalence tests, this value is crucial since it defines the interval of equivalence. Usually, this value is between 0.01 and 0.20.

Note that this value must be a positive number.

Max Equivalence Ratio

This is the largest value of the ratio of the two sensitivities that will still result in the conclusion of diagnostic equivalence. When running equivalence tests, this value is crucial since it defines the interval of equivalence. Usually, this value is between 1.05 and 2.0.

Note that this value must be greater than one.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Binary Diagnostic Test of Paired Samples

This section presents an example of how to enter data and run an analysis. In this example, a sample of 50 individuals known to have a certain disease was selected. For this study, Test 1 refers to a new, cheaper, less-invasive diagnostic test and Test 2 refers to the standard diagnostic test that is currently being used. The results are summarized into the following table:

	Test 2 Result		
Test 1 Result	Positive	Negative	Total
Positive	31	5	36
Negative	4	10	14
Total	35	15	50

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Binary Diagnostic Tests – Paired Samples window.

1 Open the Binary Diagnostic Tests – Paired Samples window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **Binary** – **Paired Samples**. The Binary Diagnostic Tests – Paired Samples procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Enter the data.

- Select the **Data tab**.
- In the **X11** box, enter **31**.
- In the **X10** box, enter **5**.
- In the **X01** box, enter **4**.
- In the **X00** box, enter **10**.

3 Set the other options.

- Set the **Difference C.I. Method** to **Score (Nam RMLE)**.
- Set the **Ratio C.I. Method** to **Score (Nam Blackwelder)**
- Set the **Max Equivalence Difference** to **0.2**.
- Set the **Max Equivalence Ratio** to **1.25**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data and Proportions

Counts

Test 1 (New) Result	Test 2 (Standard) Result Positive	Negative	Total
Positive	31	5	36
Negative	4	10	14
Total	35	15	50

Table Proportions

Test 2 (Standard) Result Positive	Negative	Total
0.6200	0.1000	0.7200
0.0800	0.2000	0.2800
0.7000	0.3000	1.0000

Row Proportions

Test 1 (New) Result	Test 2 (Standard) Result Positive	Negative	Total
Positive	0.8611	0.1389	1.0000
Negative	0.2857	0.7143	1.0000
Total	0.7000	0.3000	1.0000

Column Proportions

Test 2 (Standard) Result Positive	Negative	Total
0.8857	0.3333	0.7200
0.1143	0.6667	0.2800
1.0000	1.0000	1.0000

These reports display the counts that were entered along with various proportions that make interpreting the table easier. Note that Test 1's sensitivity of 0.7200 and Test 2's sensitivity of 0.7000 are displayed in the margins of the Table Proportions table.

Sensitivity Confidence Intervals

Statistic	Test	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Sensitivity (Se1)	1	0.7200	0.5833	0.8253
Sensitivity (Se2)	2	0.7000	0.5625	0.8090
Difference (Se1-Se2)		0.0200	-0.1094	0.1511
Ratio (Se1/Se2)		1.0286	0.8524	1.2491

Notes:
Sensitivity: proportion of those that actually have the condition for which the diagnostic test is positive.
Difference confidence limits based on Nam's RMLE method.
Ratio confidence limits based on Blackwelder and Nam's method.

This report displays the sensitivity for each test as well as corresponding confidence interval. It also shows the value and confidence interval for the difference and ratio of the sensitivity. Note that for a perfect diagnostic test, the sensitivity would be one. Hence, the larger the values the better.

Note that the type of confidence interval for the difference and ratio is specified on the Data panel.

Confidence Intervals for the Odds Ratio

Statistic	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Exact Conditional Binomial	1.2500	0.2690	6.2995
Maximum Likelihood	1.2500	0.3357	4.6549

Notes:
Odds Ratio = Odds(True Condition = +) / Odds(True Condition = -)
where
Odds(Condition) = P(Positive Test | Condition) / P(Negative Test | Condition)

This report displays estimates of the odds ratio as well as its confidence interval.

Hypothesis Tests about Sensitivity Difference

Test Name	Test Sides	Null Hypothesis (H0)	Test Statistic Value	Prob Level	Conclusion at the 5.0% Level
Nam	2	Se1-Se2=0	0.1111	0.7389	Cannot Reject H0
Nam Lower	1	Se1-Se2<=0	0.3333	0.3694	Cannot Reject H0
Nam Upper	1	Se1-Se2>=0	0.3333	0.6306	Cannot Reject H0

This report displays the results of hypothesis tests comparing the sensitivities of the two diagnostic tests using Nam's test. Note that for this test, identical test results are obtained from either the test of differences or test of ratios.

Tests of Equivalence

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Difference (Se1-Se2)	0.0051	-0.0859	0.1274	-0.2000	0.2000	Yes
Ratio (Se1/Se2)	0.0247	0.8833	1.2039	0.8000	1.2500	Yes

Notes:
Equivalence is concluded when the confidence limits fall completely inside the equivalence bounds.
Difference confidence limits based on Nam's RMLE method.
Ratio confidence limits based on Blackwelder and Nam's method.

This report displays the results of the equivalence tests of sensitivity, one based on the difference and the other based on the ratio. Equivalence is concluded if the confidence limits are inside the equivalence bounds.

Prob Level

The probability level is the smallest value of alpha that would result in rejection of the null hypothesis. It is interpreted as any other significance level. That is, reject the null hypothesis when this value is less than the desired significance level.

Note that for many types of confidence limits, a closed form solution for this value does not exist and it must be searched for.

Confidence Limits

These are the lower and upper confidence limits calculated using the method you specified. Note that for equivalence tests, these intervals use twice the alpha. Hence, for a 5% equivalence test, the confidence coefficient is 0.90, not 0.95.

Lower and Upper Bounds

These are the equivalence bounds. Values of the difference (ratio) inside these bounds are defined as being equivalent. Note that this value does not come from the data. Rather, you have to set it. These bounds are crucial to the equivalence test and they should be chosen carefully.

Reject H0 and Conclude Equivalence at the 5% Significance Level

This column gives the result of the equivalence test at the stated level of significance. Note that when you reject H0, you can conclude equivalence. However, when you do not reject H0, you cannot conclude nonequivalence. Instead, you conclude that there was not enough evidence in the study to reject the null hypothesis.

Tests Showing the Sensitivity Noninferiority of Test2 Compared to Test1

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Diff. (Se1-Se2)	0.0051	-0.0859	0.1274	-0.2000	0.2000	Yes
Ratio (Se1/Se2)	0.0247	0.8833	1.2039	0.8000	1.2500	Yes

Notes:
H0: The Sensitivity of Test2 is inferior to Test1.
Ha: The Sensitivity of Test2 is noninferior to Test1.
The noninferiority of Test2 compared to Test1 is concluded when the upper c.l. < upper bound.
Difference confidence limits based on Nam's RMLE method.
Ratio confidence limits based on Blackwelder and Nam's method.

This report displays the results of two noninferiority tests of sensitivity, one based on the difference and the other based on the ratio. Report definitions are identical with those above for equivalence.

Chapter 537

Binary Diagnostic Tests – Two Independent Samples

Introduction

An important task in diagnostic medicine is to measure the accuracy of two diagnostic tests. This can be done by comparing summary measures of diagnostic accuracy such as *sensitivity* or *specificity* using a statistical test. Often, you want to show that a new test is similar to another test, in which case you use an equivalence test. Or, you may wish to show that a new diagnostic test is not inferior to the existing test, so you use a non-inferiority test. All of these hypothesis tests are available in this procedure for the important case when the diagnostic tests provide a binary (yes or no) result.

The results of such studies can be displayed in two 2-by-2 tables in which the true condition is shown as the rows and the diagnostic test results are shown as the columns.

	Diagnostic Test 1 Result			**Diagnostic Test 2 Result**		
True Condition	Positive	Negative	Total	Positive	Negative	Total
Present (True)	*T11*	*T10*	*n11*	*T21*	*T20*	*n21*
Absent (False)	*F11*	*F10*	*n10*	*F21*	*F20*	*n20*
Total	*m11*	*m10*	*N1*	*m21*	*m20*	*N2*

Data such as this can be analyzed using standard techniques for comparing two proportions which are presented in the chapter on Two Proportions. However, specialized techniques have been developed for dealing specifically with the questions that arise from such a study. These techniques are presented in chapter 5 of the book by Zhou, Obuchowski, and McClish (2002).

Test Accuracy

Several measures of a diagnostic test's accuracy are available. Probably the most popular measures are the test's *sensitivity* and the *specificity*. Sensitivity is the proportion of those that have the condition for which the diagnostic test is positive. Specificity is the proportion of those that do not have the condition for which the diagnostic test is negative. Other accuracy measures that have been proposed are the likelihood ratio and the odds ratio. Study designs anticipate that the sensitivity and specificity of the two tests will be compared.

Comparing Sensitivity and Specificity

Suppose you arrange the results of two diagnostic tests into two 2-by-2 tables as follows:

	Standard or Reference			**Treatment or Experimental**		
	Diagnostic Test 1 Result			**Diagnostic Test 2 Result**		
True Condition	Positive	Negative	Total	Positive	Negative	Total
Present (True)	*T11*	*T10*	*n11*	*T21*	*T20*	*n21*
Absent (False)	*F11*	*F10*	*n10*	*F21*	*F20*	*n20*
Total	*m11*	*m10*	*N1*	*m21*	*m20*	*N2*

Hence, the study design include $N = N1 + N2$ patients.

The hypotheses of interest when comparing the sensitivities (Se) of two diagnostic tests are either the difference hypotheses

$$H_O: Se1 - Se2 = 0 \text{ versus } H_A: Se1 - Se2 \neq 0$$

or the ratio hypothesis

$$H_O: Se1 / Se2 = 1 \text{ versus } H_A: Se1 / Se2 \neq 1$$

Similar sets of hypotheses may be defined for the difference or ratio of the specificities (Sp) as

$$H_O: Sp1 - Sp2 = 0 \text{ versus } H_A: Sp1 - Sp2 \neq 0$$

and

$$H_O: Sp1 / Sp2 = 1 \text{ versus } H_A: Sp1 / Sp2 \neq 1$$

Note that the difference hypotheses usually require a smaller sample size for comparable statistical power, but the ratio hypotheses may be more convenient.

The sensitivities are estimated as

$$\hat{Se}1 = \frac{T11}{n11} \text{ and } \hat{Se}2 = \frac{T21}{n21}$$

and the specificities are estimated as

$$\hat{Sp}1 = \frac{F10}{n10} \text{ and } \hat{Sp}2 = \frac{F20}{n20}$$

The sensitivities and specificities of the two diagnostic test may be compared using either their difference or their ratio.

As can be seen from the above, comparison of the sensitivity or the specificity reduces to the problem of comparing two independent binomial proportions. Hence the formulas used for hypothesis testing and confidence intervals are the same as presented in the chapter on testing two independent proportions. We refer you to that chapter for further details.

Data Structure

This procedure does not use data from the database. Instead, you enter the values directly into the panel. The data are entered into two tables. The table on the left represents the existing (standard) test. The table on the right contains the data for the new diagnostic test.

Zero Cells

Although zeroes are valid values, they make direct calculation of ratios difficult. One popular technique for dealing with the difficulties of zero values is to enter a small ‘delta’ value such as 0.50 or 0.25 in the zero cells so that division by zero does not occur. Such method are controversial, but they are commonly used. Probably the safest method is to use the hypotheses in terms of the differences rather than ratios when zeroes occur, since these may be calculated without adding a delta.

Procedure Options

This section describes the options available in this procedure.

Data Tab

Enter the data values directly on this panel.

Data Values

T11 and T21

This is the number of patients that had the condition of interest and responded positively to the diagnostic test. The first character is the true condition: (T)rue or (F)alse. The second character is the test number: 1 or 2. The third character is the test result: 1=positive or 0=negative.

The value entered must be a non-negative number. Many of the reports require the value to be greater than zero.

T10 and T20

This is the number of patients that had the condition of interest but responded negatively to the diagnostic test. The first character is the true condition: (T)rue or (F)alse. The second character is the test number: 1 or 2. The third character is the test result: 1=positive or 0=negative.

The value entered must be a non-negative number. Many of the reports require the value to be greater than zero.

F11 and F21

This is the number of patients that did not have the condition of interest but responded positively to the diagnostic test. The first character is the true condition: (T)rue or (F)alse. The second character is the test number: 1 or 2. The third character is the test result: 1=positive or 0=negative.

The value entered must be a non-negative number. Many of the reports require the value to be greater than zero.

F10 and F20

This is the number of patients that did not have the condition of interest and responded positively to the diagnostic test. The first character is the true condition: (T)rue or (F)alse. The second character is the test number: 1 or 2. The third character is the test result: 1=positive or 0=negative.

The value entered must be a non-negative number. Many of the reports require the value to be greater than zero.

Confidence Interval Method

Difference C.I. Method

This option specifies the method used to calculate the confidence intervals of the proportion differences. These methods are documented in detail in the Two Proportions chapter. The recommended method is the skewness-corrected score method of Gart and Nam.

Ratio C.I. Method

This option specifies the method used to calculate the confidence intervals of the proportion ratios. These methods are documented in detail in the Two Proportions chapter. The recommended method is the skewness-corrected score method of Gart and Nam.

Report Options

Alpha - Confidence Intervals

The confidence coefficient to use for the confidence limits of the difference in proportions. 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05.

Alpha - Hypothesis Tests

This is the significance level of the hypothesis tests, including the equivalence and noninferiority tests. Typical values are between 0.01 and 0.10. The most common choice is 0.05.

Proportion Decimals

The number of digits to the right of the decimal place to display when showing proportions on the reports.

Probability Decimals

The number of digits to the right of the decimal place to display when showing probabilities on the reports.

Equivalence or Non-Inferiority Settings

Max Equivalence Difference

This is the largest value of the difference between the two proportions (sensitivity or specificity) that will still result in the conclusion of diagnostic equivalence. When running equivalence tests, this value is crucial since it defines the interval of equivalence. Usually, this value is between 0.01 and 0.20.

Note that this value must be a positive number.

Max Equivalence Ratio

This is the largest value of the ratio of the two proportions (sensitivity or specificity) that will still result in the conclusion of diagnostic equivalence. When running equivalence tests, this value is crucial since it defines the interval of equivalence. Usually, this value is between 1.05 and 2.0.

Note that this value must be greater than one.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Binary Diagnostic Test of Two Independent Samples

This section presents an example of how to enter data and run an analysis. In this example, samples of 50 individuals known to have a certain disease and 50 individuals without the disease where divided into two, equal-sized groups. Half of each group was given diagnostic test 1 and the other half was given diagnostic test 2. The results are summarized into the following tables:

	Standard or Reference Diagnostic Test 1 Result			Treatment or Experimental Diagnostic Test 2 Result		
True Condition	Positive	Negative	Total	Positive	Negative	Total
Present (True)	20	5	25	21	4	25
Absent (False)	7	18	25	5	20	25
Total	27	23	50	26	24	50

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Binary Diagnostic Tests – Two Independent Samples window.

1 Open the Binary Diagnostic Tests – Two Independent Samples window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **Binary - Two Independent Samples**. The Binary Diagnostic Tests – Two Independent Samples procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Enter the data.

- Select the **Data tab**.
- In the **T11** box, enter **20**.
- In the **T10** box, enter **5**.
- In the **F11** box, enter **7**.
- In the **F10** box, enter **18**.
- In the **T11** box, enter **21**.
- In the **T10** box, enter **4**.
- In the **F11** box, enter **5**.
- In the **F10** box, enter **20**.

3 Set the other options.

- Set the **Difference C.I. Method** to **Score w/Skewness(Gart-Nam)**.
- Set the **Ratio C.I. Method** to **Score w/Skewness(Gart-Nam)**.
- Set the **Max Equivalence Difference** to **0.25**.
- Set the **Max Equivalence Ratio** to **1.25**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data and Proportions Sections

Counts for Test 1

True Condition	Diagnostic Test Result Positive	Negative	Total
Present	20	5	25
Absent	7	18	25
Total	27	23	50

Counts for Test 2

Diagnostic Test Result Positive	Negative	Total
21	4	25
5	20	25
26	24	50

Table Proportions for Test 1

True Condition	Diagnostic Test Result Positive	Negative	Total
Present	0.4000	0.1000	0.5000
Absent	0.1400	0.3600	0.5000
Total	0.5400	0.4600	1.0000

Table Proportions for Test 2

Diagnostic Test Result Positive	Negative	Total
0.4200	0.0800	0.5000
0.1000	0.4000	0.5000
0.5200	0.4800	1.0000

Row Proportions for Test 1

True Condition	Diagnostic Test Result Positive	Negative	Total
Present	0.8000	0.2000	1.0000
Absent	0.2800	0.7200	1.0000
Total	0.5400	0.4600	1.0000

Row Proportions for Test 2

Diagnostic Test Result Positive	Negative	Total
0.8400	0.1600	1.0000
0.2000	0.8000	1.0000
0.5200	0.4800	1.0000

Column Proportions for Test 1

True Condition	Diagnostic Test Result Positive	Negative	Total
Present	0.7407	0.2174	0.5000
Absent	0.2593	0.7826	0.5000
Total	1.0000	1.0000	1.0000

Column Proportions for Test 2

Diagnostic Test Result Positive	Negative	Total
0.8077	0.1667	0.5000
0.1923	0.8333	0.5000
1.0000	1.0000	1.0000

These reports display the counts that were entered for the two tables along with various proportions that make interpreting the table easier. Note that the sensitivity and specificity are displayed in the Row Proportions table.

Sensitivity Confidence Intervals Section

Statistic	Test	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Sensitivity (Se1)	1	0.8000	0.6087	0.9114
Sensitivity (Se2)	2	0.8400	0.6535	0.9360
Difference (Se1-Se2)		-0.0400	-0.2618	0.1824
Ratio (Se1/Se2)		0.9524	0.7073	1.2659

Sensitivity: proportion of those that actually have the condition for which the diagnostic test is positive.
Difference confidence limits based on Gart and Nam's score method with skewness correction.
Ratio confidence limits based on Gart and Nam's score method with skewness correction.

This report displays the sensitivity for each test as well as corresponding confidence interval. It also shows the value and confidence interval for the difference and ratio of the sensitivity. Note that for a perfect diagnostic test, the sensitivity would be one. Hence, the larger the values the better.

Also note that the type of confidence interval for the difference and ratio is specified on the Data panel. The Wilson score method is used to calculate the individual confidence intervals for the sensitivity of each test.

Specificity Confidence Intervals Section

Statistic	Test	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Specificity (Sp1)	1	0.7200	0.5242	0.8572
Specificity (Sp2)	2	0.8000	0.6087	0.9114
Difference (Sp1-Sp2)		-0.0800	-0.3173	0.1616
Ratio (Sp1/Sp2)		0.9000	0.6326	1.2501

Notes:
Specificity: proportion of those that do not have the condition for which the diagnostic test is negative.
Difference confidence limits based on Gart and Nam's score method with skewness correction.
Ratio confidence limits based on Gart and Nam's score method with skewness correction.

This report displays the specificity for each test as well as corresponding confidence interval. It also shows the value and confidence interval for the difference and ratio of the specificity. Note that for a perfect diagnostic test, the specificity would be one. Hence, the larger the values the better.

Also note that the type of confidence interval for the difference and ratio is specified on the Data panel. The Wilson score method is used to calculate the individual confidence intervals for the specificity of each test.

Sensitivity & Specificity Hypothesis Test Section

Hypothesis Test of	Value	Chi-Square	Prob Level	Decision at 5.0% Level
Se1 = Se2	-0.0400	0.1355	0.7128	Cannot Reject H0
Sp1 = Sp2	-0.0800	0.4386	0.5078	Cannot Reject H0

This report displays the results of hypothesis tests comparing the sensitivity and specificity of the two diagnostic tests. The Pearson chi-square test statistic and associated probability level is used.

Likelihood Ratio Section

Likelihood Ratio Section

Statistic	Test	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
LR(Test=Posivitive)	1	2.8571	1.5948	6.1346
	2	4.2000	2.1020	10.8255
LR(Test=Negative)	1	0.2778	0.1064	0.5729
	2	0.2000	0.0669	0.4405

Notes:
LR(Test = +) = P(Test = + | True Condition = +) / P(Test = + | True Condition = -).
LR(Test = +) > 1 indicates a positive test is more likely among those in which True Condition = +.

LR(Test = -): P(Test = - | True Condition = +) / P(Test = - | True Condition = -).
LR(Test = -) < 1 indicates a negative test is more likely among those in which True = -.

This report displays the positive and negative likelihood ratios with their corresponding confidence limits. You would want LR(+) > 1 and LR(-) < 1, so place close attention that the lower limit of LR(+) is greater than one and that the upper limit of LR(-) is less than one.

Note the LR(+) means LR(Test=Positive). Similarly, LR(-) means LR(Test=Negative).

Odds Ratio Section

Statistic	Test	Value	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Odds Ratio (+ 1/2)	1	9.1939	2.5893	32.6452
	2	17.8081	4.4579	71.1386
Odds Ratio (Fleiss)	1	9.1939	2.3607	48.8331
	2	17.8081	4.1272	123.3154

Notes:
Odds Ratio = Odds(True Condition = +) / Odds(True Condition = -)
where
Odds(Condition) = P(Positive Test | Condition) / P(Negative Test | Condition)

This report displays estimates of the odds ratio as well as its confidence interval. Because of the better coverage probabilities of the Fleiss confidence interval, we suggest that you use this interval.

Hypothesis Tests of the Equivalence of Sensitivity

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Diff. (Se1-Se2)	0.0314	-0.2246	0.1446	-0.2500	0.2500	Yes
Ratio (Se1/Se2)	0.1109	0.7471	1.2021	0.8000	1.2500	No

Notes:
Equivalence is concluded when the confidence limits fall completely inside the equivalence bounds.
Difference confidence limits based on Gart and Nam's score method with skewness correction.
Ratio confidence limits based on Gart and Nam's score method with skewness correction.

This report displays the results of the equivalence tests of sensitivity, one based on the difference and the other based on the ratio. Equivalence is concluded if the confidence limits are inside the equivalence bounds.

Prob Level

The probability level is the smallest value of alpha that would result in rejection of the null hypothesis. It is interpreted as any other significance level. That is, reject the null hypothesis when this value is less than the desired significance level.

Note that for many types of confidence limits, a closed form solution for this value does not exist and it must be searched for.

Confidence Limits

These are the lower and upper confidence limits calculated using the method you specified. Note that for equivalence tests, these intervals use twice the alpha. Hence, for a 5% equivalence test, the confidence coefficient is 0.90, not 0.95.

Lower and Upper Bounds

These are the equivalence bounds. Values of the difference (ratio) inside these bounds are defined as being equivalent. Note that this value does not come from the data. Rather, you have to set it. These bounds are crucial to the equivalence test and they should be chosen carefully.

Reject H0 and Conclude Equivalence at the 5% Significance Level

This column gives the result of the equivalence test at the stated level of significance. Note that when you reject H0, you can conclude equivalence. However, when you do not reject H0, you cannot conclude nonequivalence. Instead, you conclude that there was not enough evidence in the study to reject the null hypothesis.

Hypothesis Tests of the Equivalence of Specificity

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Diff. (Sp1-Sp2)	0.0793	-0.2787	0.1216	-0.2500	0.2500	No
Ratio (Sp1/ Sp2)	0.2385	0.6744	1.1799	0.8000	1.2500	No

Notes:
Equivalence is concluded when the confidence limits fall completely inside the equivalence bounds.
Difference confidence limits based on Gart and Nam's score method with skewness correction.
Ratio confidence limits based on Gart and Nam's score method with skewness correction.

This report displays the results of the equivalence tests of specificity, one based on the difference and the other based on the ratio. Report definitions are identical with those above for sensitivity.

Hypothesis Tests of the Noninferiority of Sensitivity

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Noninferiority at the 5.0% Significance Level
Diff. (Se1-Se2)	0.0061	-0.2246	0.1446	-0.2500	0.2500	Yes
Ratio (Se1/Se2)	0.0297	0.7471	1.2021	0.8000	1.2500	Yes

Notes:
H0: The sensitivity of Test2 is inferior to Test1.
Ha: The sensitivity of Test2 is noninferior to Test1.
The noninferiority of Test2 compared to Test1 is concluded when the upper c.l. < upper bound.
Difference confidence limits based on Gart and Nam's score method with skewness correction.
Ratio confidence limits based on Gart and Nam's score method with skewness correction.

This report displays the results of two noninferiority tests of sensitivity, one based on the difference and the other based on the ratio. The noninferiority of test 2 as compared to test 1 is concluded if the upper confidence limit is less than the upper bound. The columns are as defined above for equivalence tests.

Hypothesis Tests of the Noninferiority of Specificity

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Noninferiority at the 5.0% Significance Level
Diff. (Sp1-Sp2)	0.0041	-0.2787	0.1216	-0.2500	0.2500	Yes
Ratio (Sp1/Sp2)	0.0250	0.6744	1.1799	0.8000	1.2500	Yes

Notes:
H0: The specificity of Test2 is inferior to Test1.
Ha: The specificity of Test2 is noninferior to Test1.
The noninferiority of Test2 compared to Test1 is concluded when the upper c.l. < upper bound.
Difference confidence limits based on Gart and Nam's score method with skewness correction.
Ratio confidence limits based on Gart and Nam's score method with skewness correction.

This report displays the results of the noninferiority tests of specificity. Report definitions are identical with those above for sensitivity.

537-12 Binary Diagnostic Tests – Two Independent Samples

Chapter 538

Binary Diagnostic Tests – Clustered Samples

Introduction

A *cluster randomization trial* occurs when whole groups or *clusters* of individuals are treated together. In the two-group case, each cluster is randomized to receive a particular treatment. In the paired case, each group receives both treatments. The unique feature of this design is that each cluster is treated the same. The usual binomial assumptions do not hold for such a design because the individuals within a cluster cannot be assumed to be independent. Examples of such clusters are clinics, hospitals, cities, schools, or neighborhoods.

When the results of a cluster randomization diagnostic trial are binary, the diagnostic accuracy of the tests is commonly summarized using the test *sensitivity* or *specificity*. Sensitivity is the proportion of those that have the condition for which the diagnostic test is positive. Specificity is the proportion of those that do not have the condition for which the diagnostic test is negative.

Often, you want to show that a new test is similar to another test, in which case you use an equivalence test. Or, you may wish to show that a new diagnostic test is not inferior to the existing test, so you use a non-inferiority test.

Specialized techniques have been developed for dealing specifically with the questions that arise from such a study. These techniques are presented in chapters 4 and 5 of the book by Zhou, Obuchowski, and McClish (2002) under the heading Clustered Binary Data. These techniques are referred to as the *ratio estimator* approach in Donner and Klar (2000).

Comparing Sensitivity and Specificity

These results apply for either an independent-group design in which each cluster receives only one diagnostic test or a paired design in which each cluster receives both diagnostic tests. The results for a particular cluster and test combination may be arranged in a 2-by-2 table as follows:

	Diagnostic Test Result		
True Condition	Positive	Negative	Total
Present (True)	*T1*	*T0*	*n1*
Absent (False)	*F1*	*F0*	*n0*
Total	*m1*	*m0*	*N*

The hypothesis set of interest when comparing the sensitivities (Se) of two diagnostic tests are

$$H_O: Se1 - Se2 = 0 \text{ versus } H_A: Se1 - Se2 \neq 0$$

A similar set of hypotheses may be defined for the difference of the specificities (Sp) as

$$H_O: Sp1 - Sp2 = 0 \text{ versus } H_A: Sp1 - Sp2 \neq 0$$

For each table, the sensitivity is estimated using

$$\hat{Se} = \frac{T1}{n1}$$

and the specificity is estimated using

$$\hat{Sp} = \frac{F0}{n0}$$

The hypothesis of equal difference in sensitivity can be tested using the following *z-test*, which follows the normal distribution approximately, especially when the number of clusters is over twenty.

$$Z_{Se} = \frac{\hat{Se}_1 - \hat{Se}_2}{\sqrt{\hat{V}_{Se1-Se2}}}$$

where

$$\hat{Se}_i = \frac{\sum_{j=1}^{K} n_{1ij}\hat{Se}_{ij}}{\sum_{j=1}^{K} n_{1ij}}$$

$$\hat{V}_{Se_1-Se_2} = \hat{Var}\left(\hat{Se}_1\right) + \hat{Var}\left(\hat{Se}_2\right) - 2\hat{Cov}\left(\hat{Se}_1, \hat{Se}_2\right)$$

$$\hat{Var}\left(\hat{Se}_i\right) = \frac{1}{K_i\left(K_i - 1\right)} \sum_{j=1}^{K_i} \left(\frac{n_{ij}}{\overline{n}_i}\right)^2 \left(\hat{Se}_{ij} - \hat{Se}_i\right)^2, \quad i = 1,2$$

$$\hat{C}ov\left(\hat{S}e_1, \hat{S}e_2\right) = \frac{1}{K(K-1)} \sum_{j=1}^{K} \left(\frac{n_{1j}}{\bar{n}}\right)^2 \left(\hat{S}e_{1j} - \bar{S}e\right)\left(\hat{S}e_{2j} - \bar{S}e\right)$$

$$\bar{S}e = \frac{\hat{S}e_1 + \hat{S}e_2}{2}$$

$$\bar{n} = \frac{1}{K} \sum_{j=1}^{K} n_{1j}$$

Here we have used K_i to represent the number of clusters receiving test *i*. For an independent design, K_1 may not be equal to K_2 and the covariance term will be zero. For a paired design, $K_1 = K_2 = K$.

Similar results may be obtained for the specificity by substituting *Sp* for *Se* in the above formulae.

Data Structure

This procedure requires four, and usually five, variables. It requires a variable containing the cluster identification, the test identification, the result identification, and the actual identification. Usually, you will add a fifth variable containing the count for the cluster, but this is not necessary if you have entered the individual data rather than the cluster data.

Here is an example of a independent-group design with four clusters per test. The Cluster column gives the cluster identification number. The Test column gives the identification number of the diagnostic test. The Result column indicates whether the result was positive (1) or negative (0). The Actual column indicates whether the disease was present (1) or absent (0). The Count column gives the number of subjects in that cluster with the indicated characteristics. Since we are dealing with 2-by-2 tables which have four cells, the data entry for each cluster requires four rows. Note that if a cell count is zero, the corresponding row may be omitted. These data are contained in the BINCLUST database.

BINCLUST dataset (subset)

Cluster	Test	Result	Actual	Count
1	1	0	0	10
1	1	1	0	3
1	1	0	1	2
1	1	1	1	21
2	1	0	0	15
2	1	1	0	2
2	1	0	1	5
2	1	1	1	10
3	1	0	0	23
3	1	1	0	3
3	1	0	1	6
3	1	1	1	31
4	1	0	0	9

Procedure Options

This section describes the options available in this procedure.

Data Tab

The options on this screen control the variables that are used in the analysis.

Cluster Variable

Cluster (Group) Variable

Specify the variable whose values indicate which cluster (group, subject, community, or hospital) is given on that row. Note that each cluster may have several rows on a database.

ID Variable

Diagnostic-Test ID Variable

Specify the variable whose values indicate which diagnostic test is recorded on this row. Note that the results of only one diagnostic test are given on each row.

Count Variable

Count Variable

This variable gives the count (frequency) for each row. Specification of this variable is optional. If it is left blank, each row will be considered to have a count of one.

Max Equivalence

Max Equivalence Difference

This is the largest value of the difference between the two proportions (sensitivity or specificity) that will still result in the conclusion of diagnostic equivalence. When running equivalence tests, this value is crucial since it defines the interval of equivalence. Usually, this value is between 0.01 and 0.20.

Note that this value must be a positive number.

Test Result Specification

Test-Result Variable

This option specifies the variable whose values indicate whether the diagnostic test was positive or negative. Thus, this variable should contain only two unique values. Often, a ‘1’ is used for positive and a ‘0’ is used for negative. The value that represents a positive value is given in the next option.

Test = Positive Value

This option specifies the value which is to be considered as a positive test result. This value must match one of the values appearing Test-Result Variable of the database. All other values will be considered to be negative. Note that the case of text data is ignored.

Condition Specification

Actual-Condition Variable

This option specifies the variable whose values indicate whether the subjects given on this row actually had the condition of interest or not. Thus, this variable should contain only two unique values. Often, a '1' is used for condition present and a '0' is used for condition absent. The value that represents a condition-present value is given in the next option.

True = Present Value

This option specifies the value which indicates that the condition is actually present. This value must match one of the values appearing Actual-Condition Variable. All other values will be considered to indicate the absence of the condition. Note that the case of text data is ignored.

Alpha

Alpha - Confidence Intervals

The confidence coefficient to use for the confidence limits of the difference in proportions. 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05.

Alpha - Hypothesis Tests

This is the significance level of the hypothesis tests, including the equivalence and noninferiority tests. Typical values are between 0.01 and 0.10. The most common choice is 0.05.

Reports Tab

The options on this screen control the appearance of the reports.

Cluster Detail Reports

Show Cluster Detail Report

Check this option to cause the detail reports to be displayed. Because of their length, you may want them omitted.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also, note that all calculations are performed in double precision regardless of which option you select here. Single precision is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

Value Labels are used to make reports more legible by assigning meaningful labels to numbers and codes.

- **Data Values**

  All data are displayed in their original format, regardless of whether a value label has been set or not.

- **Value Labels**

  All values of variables that have a value label variable designated are converted to their corresponding value label when they are output. This does not modify their value during computation.

- **Both**

  Both data value and value label are displayed.

Proportion Decimals

The number of digits to the right of the decimal place to display when showing proportions on the reports.

Probability Decimals

The number of digits to the right of the decimal place to display when showing probabilities on the reports.

Skip Line After

The names of the independent variables can be too long to fit in the space provided. If the name contains more characters than this, the rest of the output is placed on a separate line. Note: enter '1' when you want the results to be printed on two lines. Enter '100' when you want every each row's results printed on one line.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Binary Diagnostic Test of a Clustered Sample

This section presents an example of how to analyze the data contained in the BINCLUST database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Binary Diagnostic Tests – Clustered Samples window.

1 Open the BINCLUST dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **BinClust.s0**.
- Click **Open**.

2 Open the Binary Diagnostic Tests – Clustered Samples window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **Binary - Clustered Samples**. The Binary Diagnostic Tests – Clustered Samples procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Data tab**.
- Set the **Cluster (Group) Variable** to **CLUSTER**.
- Set the **Count Variable** to **COUNT**.
- Set the **Diagnostic-Test ID Variable** to **TEST**.
- Set the **Max Equivalence Difference** to **0.2**.
- Set the **Test-Result Variable** to **RESULT**.
- Set the **Test = Positive Value** to **1**.
- Set the **True = Present Value** to **1**.
- Set **Alpha - Confidence Intervals** to **0.05**.
- Set **Alpha - Hypothesis Tests** to **0.05**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Section

Parameter	Value	Parameter	Value
Cluster Variable	Cluster	Rows Scanned	32
Test Variable	Test(1, 2)	Rows Filtered	0
Actual Variable	Actual(+=1)	Rows Missing	0
Result Variable	Result(+=1)	Rows Used	32
Count Variable	Count	Clusters	8

This report records the variables that were used and the number of rows that were processed.

Sensitivity Confidence Intervals Section

Statistic	Test	Value	Standard Deviation	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Sensitivity (Se1)	1	0.8214	0.0442	0.7347	0.9081
Sensitivity (Se2)	2	0.7170	0.0518	0.6154	0.8185
Difference (Se1-Se2)		0.1044	0.0681	-0.0291	0.2380
Covariance (Se1 Se2)		0.0000			

This report displays the sensitivity for each test as well as the corresponding confidence interval. It also shows the value and confidence interval for the difference of the sensitivities. Note that for a perfect diagnostic test, the sensitivity would be one. Hence, the larger the values the better.

Specificity Confidence Intervals Section

Statistic	Test	Value	Standard Deviation	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Specificity (Sp1)	1	0.8636	0.0250	0.8147	0.9126
Specificity (Sp2)	2	0.7053	0.0768	0.5547	0.8559
Difference (Sp1-Sp2)		0.1584	0.0808	0.0000	0.3167
Covariance (Sp1 Sp2)		0.0000			

This report displays the specificity for each test as well as corresponding confidence interval. It also shows the value and confidence interval for the difference. Note that for a perfect diagnostic test, the specificity would be one. Hence, the larger the values the better.

Sensitivity & Specificity Hypothesis Test Section

Hypothesis Test of	Value	Z Value	Prob Level	Reject H0 at 5.0% Level
Se1 = Se2	0.1044	1.5334	0.1252	No
Sp1 = Sp2	0.1584	1.9602	0.0500	Yes

This report displays the results of hypothesis tests comparing the sensitivity and specificity of the two diagnostic tests. The *z* test statistic and associated probability level is used.

Hypothesis Tests of the Equivalence

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Diff. (Se1-Se2)	0.0803	-0.0076	0.2165	-0.2000	0.2000	No
Diff. (Sp1-Sp2)	0.3032	0.0255	0.2913	-0.2000	0.2000	No

Notes:
Equivalence is concluded when the confidence limits fall completely inside the equivalence bounds.

This report displays the results of the equivalence tests of sensitivity (Se1-Se2) and specificity (Sp1-Sp2), based on the difference. Equivalence is concluded if the confidence limits are inside the equivalence bounds.

Prob Level

The probability level is the smallest value of alpha that would result in rejection of the null hypothesis. It is interpreted as any other significance level. That is, reject the null hypothesis when this value is less than the desired significance level.

Confidence Limits

These are the lower and upper confidence limits calculated using the method you specified. Note that for equivalence tests, these intervals use twice the alpha. Hence, for a 5% equivalence test, the confidence coefficient is 0.90, not 0.95.

Lower and Upper Bounds

These are the equivalence bounds. Values of the difference inside these bounds are defined as being equivalent. Note that this value does not come from the data. Rather, you have to set it. These bounds are crucial to the equivalence test and they should be chosen carefully.

Reject H0 and Conclude Equivalence at the 5% Significance Level

This column gives the result of the equivalence test at the stated level of significance. Note that when you reject H0, you can conclude equivalence. However, when you do not reject H0, you cannot conclude nonequivalence. Instead, you conclude that there was not enough evidence in the study to reject the null hypothesis.

Hypothesis Tests of the Noninferiority of Test2 Compared to Test1

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Noninferiority at the 5.0% Significance Level
Diff. (Se1-Se2)	0.0803	-0.0076	0.2165	-0.2000	0.2000	No
Diff (Sp1-Sp2)	0.3032	0.0255	0.2913	-0.2000	0.2000	No

Notes:
H0: The sensitivity/specificity of Test2 is inferior to Test1.
Ha: The sensitivity/specificity of Test2 is noninferior to Test1.
The noninferiority of Test2 compared to Test1 is concluded when the upper C(0).l. < upper bound.

This report displays the results of noninferiority tests of sensitivity and specificity. The noninferiority of test 2 as compared to test 1 is concluded if the upper confidence limit is less than the upper bound. The columns are as defined above for equivalence tests.

Cluster Count Detail Section

Cluster	Test	True Pos (TP)	False Neg (FN)	True Neg (TN)	False Pos (FP)	Total True	Total False	Total Pos	Total Neg	Total
1	1	21	2	10	3	23	13	24	12	36
2	1	10	5	15	2	15	17	12	20	32
3	1	31	6	23	3	37	26	34	29	63
4	1	7	2	9	1	9	10	8	11	19
11	2	25	7	15	6	32	21	31	22	53
12	2	17	3	22	2	20	24	19	25	44
13	2	21	12	16	11	33	27	32	28	60
14	2	13	8	14	9	21	23	22	22	44
Total	1	69	15	57	9	84	66	78	72	150
Total	2	76	30	67	28	106	95	104	97	201

This report displays the counts that were given in the data. Note that each 2-by-2 table is represented on a single line of this table.

Cluster Proportion Detail Section

Cluster	Test	Sens. True Pos (TPR)	False Neg (FNR)	Spec. True Neg (TNR)	False Pos (FPR)	Total True	Total False	Total Pos	Total Neg	Prop. Cluster of Total
1	1	0.9130	0.0870	0.7692	0.2308	0.6389	0.3611	0.6667	0.3333	0.1026
2	1	0.6667	0.3333	0.8824	0.1176	0.4688	0.5313	0.3750	0.6250	0.0912
3	1	0.8378	0.1622	0.8846	0.1154	0.5873	0.4127	0.5397	0.4603	0.1795
4	1	0.7778	0.2222	0.9000	0.1000	0.4737	0.5263	0.4211	0.5789	0.0541
11	2	0.7813	0.2188	0.7143	0.2857	0.6038	0.3962	0.5849	0.4151	0.1510
12	2	0.8500	0.1500	0.9167	0.0833	0.4545	0.5455	0.4318	0.5682	0.1254
13	2	0.6364	0.3636	0.5926	0.4074	0.5500	0.4500	0.5333	0.4667	0.1709
14	2	0.6190	0.3810	0.6087	0.3913	0.4773	0.5227	0.5000	0.5000	0.1254
Total	1	0.8214	0.1786	0.8636	0.1364	0.5600	0.4400	0.5200	0.4800	0.4274
Total	2	0.7170	0.2830	0.7053	0.2947	0.5274	0.4726	0.5174	0.4826	0.5726

This report displays the proportions that were found in the data. Note that each 2-by-2 table is represented on a single line of this table.

Example 2 – Paired Design

Zhou (2002) presents a study of 21 subjects to compare the specificities of PET and SPECT for the diagnosis of hyperparathyroidism. Each subject had from 1 to 4 parathyroid glands that were disease-free. Only disease-free glands are needed to estimate the specificity.

The data from this study have been entered in the PET database. Open this database now. You will see that we have entered four rows for each subject. The first two rows are for the PET test (Test = 1) and the last two rows are for the SPECT test (Test = 2). Note that we have entered zero counts in several cases when necessary. During the analysis, the rows with zero counts are ignored.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Binary Diagnostic Tests – Clustered Samples window.

1 Open the PET dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **PET.s0**.
- Click **Open**.

2 Open the Binary Diagnostic Tests - Clustered Samples window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **Binary - Clustered Samples**. The Binary Diagnostic Test - Clustered Samples procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Data tab**.
- Set the **Cluster (Group) Variable** to **SUBJECT**.
- Set the **Count Variable** to **COUNT**.
- Set the **Diagnostic-Test ID Variable** to **TEST**.
- Set the **Max Equivalence Difference** to **0.2**.
- Set the **Test-Result Variable** to **RESULT**.
- Set the **Test = Positive Value** to **1**.
- Set the **True = Present Value** to **1**.
- Set **Alpha – Confidence Intervals** to **0.05**.
- Set **Alpha - Hypothesis Tests** to **0.05**.

4 4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Output

Run Summary Section

Parameter	Value	Parameter	Value
Cluster Variable	Subject	Rows Scanned	84
Test Variable	Test(1, 2)	Rows Filtered	0
Actual Variable	Actual(+=1)	Rows Missing	0
Result Variable	Result(+=1)	Rows Used	53
Count Variable	Count	Clusters	21

Specificity Confidence Intervals Section

Statistic	Test	Value	Standard Deviation	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
Specificity (Sp1)	1	0.7843	0.0696	0.6479	0.9207
Specificity (Sp2)	2	0.9020	0.0380	0.8275	0.9764
Difference (Sp1-Sp2)		-0.1176	0.0665	-0.2479	0.0127
Covariance (Sp1 Sp2)		0.0009			

Sensitivity & Specificity Hypothesis Test Section

Hypothesis Test of	Value	Z Value	Prob Level	Reject H0 at 5.0% Level
Sp1 = Sp2	-0.1176	-1.7696	0.0768	No

Hypothesis Tests of Equivalence

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Diff. (Sp1-Sp2)	0.1077	-0.2270	-0.0083	-0.2000	0.2000	No

Tests Showing the Noninferiority of Test2 Compared to Test1

Statistic	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Noninferiority at the 5.0% Significance Level
Diff (Sp1-Sp2)	0.0000	-0.2270	-0.0083	-0.2000	0.2000	Yes

This report gives the analysis of the study comparing PET (Test=1) with SPECT (Test=2). We have removed the results for sensitivity since these were not part of this database. The results show that the two specificities are not significantly different. The equivalence test shows that although the hypothesis of equality could not be rejected, the hypothesis of equivalence could not be concluded either.

Chapter 545

ROC Curves

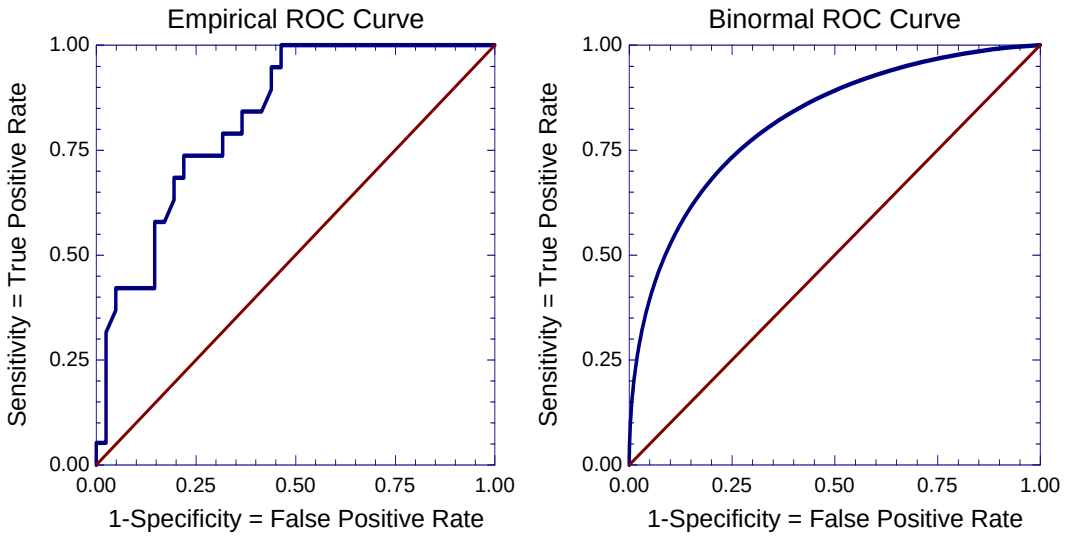
Introduction

This procedure generates both binormal and empirical (nonparametric) ROC curves. It computes comparative measures such as the whole, and partial, area under the ROC curve. It provides statistical tests comparing the AUCs and partial AUCs for paired and independent sample designs.

Discussion

A diagnostic test yields a measurement (*criterion* value) that is used to diagnose some condition of interest such as a disease. (In the sequel, we will often call the 'condition of interest' the 'disease.') The measurement might be a rating along a discrete scale or a value along a continuous scale. A positive or negative diagnosis is made by comparing the measurement to a cutoff value. If the measurement is less (or greater as the case may be) than the cutoff, the test is negative. Otherwise, the test is positive. Thus the cutoff value helps determine the rates of false positives and false negatives.

A receiver operating characteristic (ROC) curve shows the characteristics of a diagnostic test by graphing the false-positive rate (1-*specificity*) on the horizontal axis and the true-positive rate (*sensitivity*) on the vertical axis for various cutoff values. Examples of an empirical ROC curve and a binormal ROC curve are shown below.



Each point on the ROC curve represents a different cutoff value. Cutoff values that result in low false-positive rates tend to result low true-positive rates as well. As the true-positive rate increases, so does the false positive rate. Obviously, a useful diagnostic test should have a cutoff value at which the true-positive rate is high and the false-positive rate is low. In fact, a near-perfect diagnostic test would have an ROC curve that is almost vertical from (0,0) to (0,1) and then horizontal to (1,1). The diagonal line serves as a reference line since it is the ROC curve of a diagnostic test that is useless in determining the disease.

Complete discussions about ROC curves can be found in Altman (1991), Swets (1996), and Zhou et al (2002). Gehlbach (1988) provides an example of its use.

Methods for Creating ROC Curves

Several methods have been proposed to generate ROC curves. These include the binormal and the empirical (nonparametric) methods.

Binormal

The most commonly used method to generate smooth ROC curves is the binormal method popularized by a group of researchers including Metz (1978) (who developed the popular ROCFIT software). This method considers two populations: those with, and those without, the disease. It assumes that the criterion variable (or a scale-transformation of it) follows a normal distribution in each population. Using this normality assumption, a smooth ROC curve can be drawn using the sample means and variances of the two populations. Researchers have shown through various simulation studies that this *binormal* assumption is not as limiting as at first thought since non-normal data can often be transformed to a near-normal scale. **So, if you want to use this method, you should make sure that your data has been transformed so that it is nearly normal.**

Empirical or Nonparametric

An empirical (nonparametric) approach that does not depend on the normality assumptions was developed by DeLong, DeLong, and Clarke-Pearson (1988). These ROC curves are especially useful when the diagnostic test results in a continuous criterion variable.

Types of ROC Experimental Designs

Either of two experimental designs are usually employed when comparing ROC curves. These designs are *paired* or *independent samples*. Separate methods of analysis are needed to compare ROC curves depending upon which experimental design was used.

Independent Sample (Non-correlated) Designs

In this design, individuals with, and without, the disease are randomly assigned into two (or more) groups. The first group receives diagnostic test A and the second group receives diagnostic test B. Each individual receives only one diagnostic test.

Paired (Correlated) Designs

In this design, individuals with, and without, the disease each receive both diagnostic tests. This allows each subject to 'serve as their own control.'

An Example Using a Paired Design

ROC curves are explained with an example paraphrased from Gehlbach (1988). Forty-five patients with fever, headache, and a history of tick bite were classified into two groups: those with Rocky Mountain Spotted Fever (RMSF) and those without it. The serum-sodium level of each patient is measured using two techniques. We want to determine if serum-sodium level is useful in detecting RMSF, which technique is most accurate in diagnosing RMSF, and what the diagnostic cutoff value of the selected test should be. The data are presented next.

RMSF=Yes

ID	Method1	Method2	Diagnosis
1	124	122	1
2	125	124	1
3	126	125	1
4	126	125	1
5	127	126	1
6	128	126	1
7	128	127	1
8	128	128	1
9	128	128	1
10	129	128	1
11	129	130	1
12	131	130	1
13	132	133	1
14	133	133	1
15	133	134	1
16	135	134	1
17	135	134	1
18	135	134	1
19	136	136	1
20	138	138	1
21	139	140	1

RMSF=No

ID	Method1	Method2	Diagnosis
22	129	124	0
23	131	128	0
24	131	130	0
25	134	133	0
26	134	133	0
27	135	133	0
28	136	134	0
29	136	134	0
30	136	134	0
31	137	134	0
32	137	136	0
33	138	136	0
34	138	137	0
35	139	138	0
36	139	138	0
37	139	140	0
38	139	140	0
39	140	141	0
40	140	141	0
41	141	142	0
42	142	142	0
43	142	142	0
44	142	142	0
45	143	144	0

The first step in analyzing these data is to create a two-by-two table showing the diagnostic accuracy of each method at given cutoff value. If the cutoff is set at X, the table would appear as follows:

Generic Table for Cutoff=X

Method Cutoff=X	RMSF = Yes	RMSF = No
Sodium <= X, Positive	A	B
Sodium > X, Negative	C	D

The letters A, B, C, and D represent counts of the number of individuals in each of the four possible categories.

For example, if a cutoff of 130 is used to diagnose those with the disease, the tables for each sodium measurement method would be:

Table for Method 1, Cutoff = 130

Method 1 Cutoff=130	**RMSF=Yes**	**RMSF=No**
Sodium<=130, Positive	**11**	**1**
Sodium>130, Negative	**10**	**23**

Table for Method 2, Cutoff = 130

Method 2 Cutoff=130	RMSF=Yes	RMSF=No
Sodium<=130, **Positive**	12	3
Sodium>130, **Negative**	9	21

If a cutoff of 137 is used to diagnose those with the disease, the tables for each sodium measurement method would be:

Table for Method 1, Cutoff = 137

Method 1 Cutoff=137	RMSF=Yes	RMSF=No
Sodium<=137, **Positive**	19	11
Sodium>137, **Negative**	2	13

Table for Method 2, Cutoff = 137

Method 2 Cutoff=137	RMSF=Yes	RMSF=No
Sodium<=137, **Positive**	19	13
Sodium>137, **Negative**	12	11

As you study these tables, you can see changing the cutoff value changes the table counts. An ROC curve is constructed by creating many of these tables and plotting the sensitivity versus one minus the specificity.

Definition of Terms

We will now define the indices that are used to create ROC curves.

Sensitivity

Sensitivity is the proportion of those with the disease that are correctly identified as having the disease by the test. In terms of our two-by-two tables, sensitivity = *A/(A+C)*.

Specificity

Specificity is the proportion of those without the disease that are correctly identified as not having the disease by the test. In terms of our two-by-two tables, specificity = *D/(B+D)*.

Prevalence

Prevalence is the overall proportion of individuals with the disease. In terms of our two-by-two tables, prevalence = *(A+C)/(A+B+C+D)*. Notice that since the prevalence is defined in terms of the marginal values, it does not depend on the cutoff value.

Positive Predictive Value (PPV)

PPV is the proportion of individuals with positive test results who have the disease. In terms of our two-by-two tables, PPV = *(A)/(A+B)*.

Negative Predictive Value (NPV)

NPV is the proportion of individuals with negative test results who do not have the disease. In terms of our two-by-two tables, NPV = *(D)/(C+D)*.

Discussion about PPV and NPV

A problem with sensitivity and specificity is that they do not assess the probability of making a correct diagnosis. To overcome this, practitioners have developed two other indices: PPV and NPV. Unfortunately, these indices have the disadvantage that they are directly impacted by the prevalence of the disease in the population. For example, if your sampling procedure is constructed to obtain more individuals with the disease than is the case in the whole population of interest, the PPV and NPV need to be adjusted.

Using Bayes theorem, adjusted values of PPV and NPV are calculated based on new prevalence values as follows:

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specificity) \times (1 - prevalence)}$$

$$NPV = \frac{specificity \times (1 - prevalence)}{(1 - sensitivity) \times prevalence + specificity \times (1 - prevalence)}$$

Another way of interpreting these terms is as follows. The prevalence of a disease is the prior probability that a subject has the disease before the diagnostic test is run. The values of PPV and 1-NPV are the posterior probabilities of a subject having the disease after the diagnostic test is conducted.

Likelihood Ratio

The likelihood ratio statistic measures the value of the test for increasing certainty about a positive diagnosis. It is calculated as follows:

$$LR = \frac{\Pr(\textit{positive test}|\textit{disease})}{\Pr(\textit{positive test}|\textit{no disease})} = \frac{\textit{sensitivity}}{1-\textit{specificity}}$$

Finding the Optimal Criterion Value

The optimal criterion value is that value that minimizes the average cost. The approach we use was given by Metz (1978) and Zhou et al. (2002). This approach is based on an analysis of the costs (and benefits) of the four possible outcomes of a diagnostic test: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The cost of each of these outcomes must be determined. This is no small task. In fact, a whole field of study has arisen to determine these costs. Once these costs are found, the average overall cost *C* of performing a test is given by

$$C = C_0 + C_{TP}\mathrm{P}(TP) + C_{TN}\mathrm{P}(TN) + C_{FP}\mathrm{P}(FP) + C_{FN}\mathrm{P}(FN)$$

Here, C_0 is the fixed cost of performing the test, C_{TP} is the cost associated with a true positive, P(TP) is the proportion of TP's in the population, and so on. Note that P(TP) is equal to

$$\mathrm{P(TP)} = \textit{Sensitivity}\ [\mathrm{P}(\textit{Condition} = \mathrm{True})]$$

Metz (1978) showed that the point along the ROC curve where the average cost is minimum is the point where *sensitivity - m(1 - specificity)* is maximized, where

$$m = \frac{\mathrm{P}(\textit{Condition} = \mathrm{False})}{\mathrm{P}(\textit{Condition} = \mathrm{True})}\left(\frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}\right)$$

P(*Condition* = True) is called the *prevalence* of the disease. Depending on the method used to obtain the sample, it may or may not be estimated from the sample. Note that the costs enter this equation as the ratio of the net cost for a test of an individual without the disease to the net cost for a test of an individual with the disease.

Using the above result, the cut optimum cutoff may be found by scanning a report that shows *C*(*Cutoff*) for every value of the cutoff variable.

Area Under the ROC Curve (AUC)

The AUC of a Single ROC Curve

The area under an ROC curve (AUC) is a popular measure of the accuracy of a diagnostic test. Other things being equal, the larger the AUC, the better the test is a predicted the existence of the disease. The possible values of AUC range from 0.5 (no diagnostic ability) to 1.0 (perfect diagnostic ability).

The AUC has a physical interpretation. The AUC is the probability that the criterion value of an individual drawn at random from the population with the disease is larger than the criterion value of another individual drawn at random from the population without the disease.

A statistical test of usefulness of a diagnostic test is to compare it to the value 0.5. Such a statistical test can be made if we are willing to assume that the sample is large enough so that the estimated AUC follows the normal distribution. The statistical test is

$$z = \frac{\tilde{A} - 0.5}{\sqrt{\mathrm{V}\left(\tilde{A}\right)}}$$

where $\tilde{A}$ is the estimated AUC and $\mathrm{V}\left(\tilde{A}\right)$ is the estimated variance of $\tilde{A}$.

Two methods are commonly used to estimate the AUC. The first is the *binormal* method presented by Metz (1978) and McClish (1989). This method results in a smooth ROC curve from which both the complete and partial AUC may be calculated. The second method is the empirical (nonparametric) method by DeLong et al (1988). This method has become popular because it does not make the strong normality assumptions that the binormal method makes. The above *z* test may be used for both methods, as long as an appropriate estimate of $\mathrm{V}\left(\tilde{A}\right)$ is used.

The AUC of a Single Binormal ROC Curve

The formulas that we use here come from McClish (1989). Suppose there are two populations, one made up of individuals with the disease and the other made up of individuals without the disease. Further suppose that the value of a criterion variable is available for all individuals. Let *X* refer to the value of the criterion variable in the non-diseased population and *Y* refer to the value of the criterion variable in the diseased population. The binormal model assumes that both *X* and *Y* are normally distributed with different means and variances. That is,

$$X \sim N\left(\mu_x, \sigma_x^2\right),\ Y \sim N\left(\mu_y, \sigma_y^2\right)$$

The ROC curve is traced out by the function

$$\left\{FP(c), TP(c)\right\} = \left\{\Phi\left(\frac{\mu_x - c}{\sigma_x}\right), \Phi\left(\frac{\mu_y - c}{\sigma_y}\right)\right\}, \quad -\infty < c < \infty$$

where $\Phi(z)$ is the cumulative normal distribution function.

The area under the whole ROC curve is

$$\begin{aligned} A &= \int_{-\infty}^{\infty} TP(c) FP'(c)\, \mathrm{d}c \\ &= \int_{-\infty}^{\infty} \left[\Phi\left(\frac{\mu_y - c}{\sigma_y}\right)\phi\left(\frac{\mu_x - c}{\sigma_x}\right)\right] \mathrm{d}c \\ &= \Phi\left[\frac{a}{\sqrt{1+b^2}}\right] \end{aligned}$$

where

$$a = \frac{\mu_y - \mu_x}{\sigma_y} = \frac{\Delta}{\sigma_y},\ b = \frac{\sigma_x}{\sigma_y},\ \Delta = \mu_y - \mu_x$$

The area under a portion of the AUC curve is given by

$$A = \int_{c_1}^{c_2} TP(c)FP'(c)\,\mathrm{d}c$$

$$= \frac{1}{\sigma_x}\int_{c_2}^{c_1}\left[\Phi\left(\frac{\mu_y - c}{\sigma_y}\right)\phi\left(\frac{\mu_x - c}{\sigma_x}\right)\right]\mathrm{d}c$$

The partial area under an ROC curve is usually defined in terms of a range of false-positive rates rather than the criterion limits c_1 and c_2. However, the one-to-one relationship between these two quantities, given by

$$c_i = \mu_x + \sigma_x \Phi^{-1}(FP_i)$$

allows the criterion limits to be calculated from desired false-positive rates.

The MLE of *A* is found by substituting the MLE's of the means and variances into the above expression and using numerical integration. When the area under the whole curve is desired, these formulas reduce to

$$\hat{A} = \Phi\left[\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right]$$

Note that for ease of reading we will often omit the use of the *hat* to indicate an MLE in the sequel.

The variance of $\hat{A}$ is derived using the method of differentials as

$$\mathrm{V}\left(\hat{A}\right) = \left(\frac{\partial A}{\partial \Delta}\right)^2 \mathrm{V}\left(\hat{\Delta}\right) + \left(\frac{\partial A}{\partial \sigma_x^2}\right)^2 \mathrm{V}\left(s_x^2\right) + \left(\frac{\partial A}{\partial \sigma_y^2}\right)^2 \mathrm{V}\left(s_y^2\right)$$

where

$$\frac{\partial A}{\partial \Delta} = \frac{E}{\sqrt{2\pi\left(1+b^2\right)\sigma_y^2}}\left[\Phi\left(\tilde{c}_1\right) - \Phi\left(\tilde{c}_0\right)\right]$$

$$\frac{\partial A}{\partial \sigma_x^2} = \frac{E}{4\pi\left(1+b^2\right)\sigma_x\sigma_y}\left[e^{-k_0} - e^{-k_1}\right] - \frac{abE}{2\sigma_x\sigma_y\sqrt{2\pi}\left(1+b^2\right)^{3/2}}\left[\Phi\left(\tilde{c}_1\right) - \Phi\left(\tilde{c}_0\right)\right]$$

$$E = \exp\left(-\frac{a^2}{2\left(1+b^2\right)}\right)$$

$$\frac{\partial A}{\partial \sigma_y^2} = -\frac{a}{2\sigma_y}\left(\frac{\partial A}{\partial \Delta}\right) - b^2\left(\frac{\partial A}{\partial \sigma_x^2}\right)$$

$$\tilde{c}_i = \left[\Phi^{-1}\left(FP_i\right) + \frac{ab}{\left(1+b^2\right)}\right]\sqrt{\left(1+b^2\right)}$$

$$k_i = \frac{\tilde{c}_i^2}{2}$$

$$\mathrm{V}\left(\hat{\Delta}\right) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

$$\mathrm{V}\left(s_x^2\right) = \frac{2\sigma_x^4}{n_x - 1}$$

$$\mathrm{V}\left(s_y^2\right) = \frac{2\sigma_y^4}{n_y - 1}$$

Once estimates of $\hat{A}$ and $\mathrm{V}\left(\hat{A}\right)$ are calculated, hypothesis tests and confidence intervals can be calculated using standard methods. However, following the advice of Zhou et al. (2002) page 125, we use the following transformation which results in statistics that are closer to normality and ensures confidence limits that are outside the zero-one range. The transformation is

$$\hat{\psi} = \tfrac{1}{2}\ln\left(\frac{1+A}{1-A}\right)$$

The variance of $\hat{\psi}$ is estimated using

$$\mathrm{V}\left(\hat{\psi}\right) = \frac{4}{\left(1-\hat{A}^2\right)^2}\mathrm{V}\left(\hat{A}\right)$$

An $100(1-\alpha)\%$ confidence interval for ψ may then be constructed as

$$L,U = \hat{\psi} \mp z_{1-\alpha/2}\sqrt{\mathrm{V}\left(\hat{\psi}\right)}$$

Using the inverse transformation, the confidence interval for *A* is given by the two limits

$$\frac{1-e^{-L}}{1+e^{-L}} \quad \text{and} \quad \frac{1-e^{-U}}{1+e^{-U}}$$

The AUC of a Single Empirical ROC Curve

The empirical (nonparametric) method by DeLong et al (1988) is a popular method for computing the AUC. This method has become popular because it does not make the strong normality assumptions that the binormal method makes. The formula for computing this estimate of the AUC and its variance are given later in the section on comparing two empirical ROC curves.

Comparing the AUC of Two ROC Curves

Occasionally, it is of interest to compare the areas under the ROC curve (AUC) of two diagnostic tests using a hypothesis test. This may be done using either the binormal model results shown in McClish (1989) or the empirical (nonparametric) results of DeLong (1988).

Comparing the AUC of Two Empirical ROC Curves

A statistical test may be constructed that uses empirical estimates of the AUCs, their variances, and covariance. The variance and covariance formulas used depend on whether the design is paired or independent samples. Following Zhou et al. (2002) page 185, the formula to compare two AUCs is the following *z* test (which asymptotically follows the standard normal distribution) is given by

$$z = \frac{A_1 - A_2}{\sqrt{\mathrm{V}(A_1 - A_2)}}$$

where

$$\mathrm{V}(A_1 - A_2) = \mathrm{V}(A_1) + \mathrm{V}(A_2) - 2\mathrm{Cov}(A_1, A_2)$$

Independent Samples

For independent samples in which each subject receives only one of the two diagnostic tests, the covariance is zero and the two variances are

$$\mathrm{V}(A_k) = \frac{S_{T_{k1}}}{n_{k1}} + \frac{S_{T_{k0}}}{n_{k0}}$$

where

$$S_{T_{ki}} = \frac{1}{n_{ki} - 1}\sum_{j=1}^{n_{ki}}\left[\mathrm{V}(T_{kij}) - A_k\right]^2, \quad k = 1,2 \quad i = 0,1$$

$$\mathrm{V}(T_{k1i}) = \frac{1}{n_{k0} - 1}\sum_{j=1}^{n_{k0}}\psi(T_{k1i}, T_{k0j}), \quad k = 1,2$$

$$\mathrm{V}(T_{k0j}) = \frac{1}{n_{k1} - 1}\sum_{i=1}^{n_{k1}}\psi(T_{k1i}, T_{k0j}), \quad k = 1,2$$

$$A_k = \frac{\sum_{i=1}^{n_{k1}}\mathrm{V}(T_{k1i})}{n_{k1}} = \frac{\sum_{j=1}^{n_{k0}}\mathrm{V}(T_{k0j})}{n_{k0}}, \quad k = 1,2$$

$$\psi(X,Y) = \begin{cases} 0 \; if \; Y > X \\ \frac{1}{2} \; if \; Y = X \\ 1 \; if \; Y < X \end{cases}$$

Here T_{k0j} represents the observed diagnostic test result for the jth subject in group k without the disease and T_{k1j} represents the observed diagnostic test result for the jth subject in group k with the disease.

Paired Samples

For paired samples in which each subject receives both of the two diagnostic tests, the variances are given as above and the covariance is given by

$$\text{Cov}\left(A_1, A_2\right) = \frac{S_{T_{11}T_{21}}}{n_1} + \frac{S_{T_{10}T_{20}}}{n_0}$$

where

$$S_{T_{11}T_{21}} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left[\text{V}\left(T_{11j}\right) - A_1\right]\left[\text{V}\left(T_{21j}\right) - A_2\right]$$

$$S_{T_{10}T_{20}} = \frac{1}{n_0 - 1}\sum_{j=1}^{n_1}\left[\text{V}\left(T_{10j}\right) - A_1\right]\left[\text{V}\left(T_{20j}\right) - A_2\right]$$

Comparing Two Binormal AUCs

When the binormal assumption is viable, the hypothesis that the areas under the two ROC curves are equal may be tested using

$$z = \frac{A_1 - A_2}{\sqrt{\text{V}\left(A_1 - A_2\right)}}$$

Independent Samples Design

When an independent sample design is used, the variance of the difference in AUC's is the sum of the variances since the covariance is zero. That is,

$$\text{V}\left(A_1 - A_2\right) = \text{V}\left(A_1\right) + \text{V}\left(A_2\right)$$

where $\text{V}\left(A_1\right)$ and $\text{V}\left(A_2\right)$ are calculated using the formula (with obvious substitution) for $\text{V}\left(A\right)$ given above in the section on a single binormal ROC curve.

Paired Design

When a paired design is used, the variance of the difference in AUC's is

$$\text{V}\left(A_1 - A_2\right) = \text{V}\left(A_1\right) + \text{V}\left(A_2\right) - 2\text{Cov}\left(A_1, A_2\right)$$

where $\text{V}\left(A_1\right)$ and $\text{V}\left(A_2\right)$ are calculated using the formula for $\text{V}\left(A\right)$ given above in the section on a single binormal ROC curve. Since the data are paired, a covariance term must also be calculated. This is done using the differential method as follows

$$\begin{aligned}\mathrm{Cov}\left(A_1, A_2\right) = &\left(\frac{\partial A_1}{\partial \Delta_1}\right)\left(\frac{\partial A_2}{\partial \Delta_2}\right)\mathrm{Cov}\left(\hat{\Delta}_1, \hat{\Delta}_2\right) \\ &+\left(\frac{\partial A_1}{\partial \sigma^2_{x_1}}\right)\left(\frac{\partial A_2}{\partial \sigma^2_{x_2}}\right)\mathrm{Cov}\left(s^2_{x_1}, s^2_{x_2}\right) \\ &+\left(\frac{\partial A_1}{\partial \sigma^2_{y_1}}\right)\left(\frac{\partial A_2}{\partial \sigma^2_{y_2}}\right)\mathrm{Cov}\left(s^2_{y_1}, s^2_{y_2}\right)\end{aligned}$$

where

$$\mathrm{Cov}\left(\hat{\Delta}_1, \hat{\Delta}_2\right) = \frac{\rho_x \sigma_{x_1}\sigma_{x_2}}{n_x} + \frac{\rho_y \sigma_{y_1}\sigma_{y_2}}{n_y}$$

$$\mathrm{Cov}\left(s^2_{x_1}, s^2_{x_2}\right) = \frac{2\rho_x \sigma^2_{x_1}\sigma^2_{x_2}}{n_x - 1}$$

$$\mathrm{Cov}\left(s^2_{y_1}, s^2_{y_2}\right) = \frac{2\rho_y \sigma^2_{y_1}\sigma^2_{y_2}}{n_y - 1}$$

and $\rho_y\left(\rho_x\right)$ is the correlation between the two sets of criterion values in the diseased (non-diseased) population.

Transformation Achieve Normality

McClish (1989) ran simulations to study the accuracy of the normality approximation of the above z statistic for various portions of the AUC curve. She found that a logistic-type transformation resulted in a z statistic that was closer to normality. This transformation is

$$\theta\left(A\right) = \ln\left(\frac{FP_2 - FP_1 + A}{FP_2 - FP_1 - A}\right)$$

which has the inverse version

$$A = \left(FP_2 - FP_1\right)\frac{e^{\theta} - 1}{e^{\theta} + 1}$$

The variance of this quantity is given by

$$\mathrm{V}\left(\theta\right) = \left(\frac{2\left(FP_2 - FP_1\right)}{\left(FP_2 - FP_1\right)^2 - A^2}\right)^2 \mathrm{V}\left(A\right)$$

and the covariance is given by

$$\mathrm{Cov}\left(\theta_1, \theta_2\right) = \frac{4\left(FP_2 - FP_1\right)^2}{\left[\left(FP_2 - FP_1\right)^2 - A_1^2\right]\left[\left(FP_2 - FP_1\right)^2 - A_2^2\right]}\mathrm{Cov}\left(A_1, A_2\right)$$

The adjusted *z* statistic is

$$z = \frac{\theta_1 - \theta_2}{\sqrt{\mathrm{V}(\theta_1 - \theta_2)}}$$
$$= \frac{\theta_1 - \theta_2}{\sqrt{\mathrm{V}(\theta_1) + \mathrm{V}(\theta_2) - 2Cov(\theta_1, \theta_2)}}$$

Data Structure

The data are entered in two or more variables. One variable specifies the true condition of the individual. The other variable(s) contain the criterion value(s) for the tests being compared.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the analysis.

Actual Condition (Disease)

Actual Condition Variable

A binary response variable which represents whether or not the individual actually has the condition of interest. The value representing a yes is specified in the Positive Condition Value box. The values may be text or numeric.

Positive Condition Value

This is the value of the Actual Condition Variable that indicates that the individual has the condition of interest. All other values are considered as not having the condition of interest. Often, the positive value is set to '1' and the negative value is set to '0.' However, any numbering scheme may be used.

Actual Condition Prevalence

This option specifies the prevalence of the disease which is the proportion of individuals in the population that have the disease. As a proportion, this number varies between zero and one. Often an accurate estimate of this value cannot be calculated from the data, so it must be entered. Using this value, adjusted values of PPV and NPV are calculated.

Cost Benefit Ratios

This is the ratio of the net cost when the condition is absent to the net cost when the condition is present. In symbols, this is

[Cost(FP)-Benefit(TN)] / [(Cost(FN)-Benefit(TP)]

This value is used to compute the optimum criterion value. Since it is difficult to calculate this value exactly, you can enter a set of up to 4 values. These values will be used in the Cost-Benefit Report.

You can enter a list of values such as ‘0.5 0.8 0.9 1.0’ or use the special list format: ‘0.5:0.8(0.1)’.

Criterion

Criterion Variable(s)

A list of one or more criterion (test, score, discriminant, etc.) variables. If more than one variable is listed, a separate curve is drawn for each.

Note that for a paired design, both criterion variables would be specified here and the Group Variable option would be left blank. However, for an independent sample design, a single criterion variable would be specified here and a Group Variable would be specified.

Test Direction

This option specifies whether low or high values of the criterion variable indicate that the test is positive for the disease.

- **Low X = Positive**

  A low value of the criterion variable indicates a positive test result. That is, a low value will indicate a positive test.

- **High X = Positive**

  A high value of the criterion variable indicates a positive test result. That is, a high value will indicate a positive test.

Criterion List

Specify the specific values of the criterion variable to be shown on the reports. Enter ‘Data’ when you want the unique criterion values from the database to be used.

You can enter a list using numbers separated by blanks such as ‘1 2 3 4 5’ or you can use the ‘xx TO yy BY inc’ syntax or the ‘xx:yy(inc)’ syntax.

For example, entering ‘1 TO 10 BY 3’ or ‘1:10(3)’ is the same as entering ‘1 4 7 10’.

Other Variables

Frequency Variable

An optional variable containing a set of counts (frequencies). Normally, each row represents one individual. On occasion, however, each row of data may represent more than one individual. This variable contains the number of individuals that a row represents.

Group Variable

This optional variable may be used to divide the subjects into groups. It is only used when you have an independent samples design and there is just one Criterion Variable specified. If more than one Criterion Variables are specified, this variable is ignored.

When specified and used, a separate ROC curve is generated for each unique value of this variable.

Max Equivalence

Max Equivalence Difference

This value is used by the equivalence and noninferiority tests. This is the largest value of the difference in AUCs that will still result in the conclusion of equivalence of the two diagnostic tests. That is, if the true difference in AUCs is less than this amount, the two tests are assumed to be equivalent. Care must be used to be certain a reasonable value is selected.

Note that the range of equivalence is from -D to D, where 'D' is the value specified here.

We recommend a value of 0.05 as a reasonable choice. The value is usually between 0 and 0.2.

AUC Limits (Binormal Reports Only)

Upper and Lower AUC Limits

Note that these options are used with binormal reports only. They are not used with the empirical (nonparametric) reports.

The horizontal axis of the ROC curve is the proportion of false-positives (FP). Usually, the AUC is computed for full range of false-positive, i.e. from FP = 0 to FP = 1. These options let you compute the area for only the portion of FP values between these two limits. This is useful for situations in which only a portion of the ROC curve is of interest.

Note the these values must be between zero and one and that the Lower Limit must be less than the Upper Limit.

Reports Tab

The following options control the reports that are displayed.

Select Reports

ROC Data Report - Predictive Value Report

Check to display the corresponding report or plot.

Area Under Curve (AUC) Analysis: One Curve

Check to display all reports about a single AUC.

Area Under Curve (AUC) Analysis: Compare Two Curves

Check to display all reports comparing two AUCs.

Area Under Curve (AUC) Analysis: Two Curve Equivalence

Check to display all reports concerning the testing of equivalence and noninferiority.

Select ROC Plots

Empirical and Binormal

Check to display the empirical and binormal ROC plots.

Alpha

Confidence Intervals

This option specifies the value of alpha to be used in all confidence intervals. The quantity (1-Alpha) is the confidence coefficient (or confidence level) of all confidence intervals. Sets of 100 x (1 - alpha)% confidence limits are calculated. This must be a value between 0.0 and 0.5. The most common value is 0.05.

Hypothesis Tests

This option specifies the value of alpha to be used in all hypothesis tests including tests of noninferiority and equivalence. Alpha is the probability of rejecting the null hypothesis when it is true. This must be a value between 0.0 and 0.5. The most common value is 0.05.

Report Options

Skip Line After

When writing a row of information to a report, some variable names/labels may be too long to fit in the space allocated. If the name(or label) contains more characters than this, the rest of the output for that line is moved down to the next line. Most reports are designed to hold a label of up to 15 characters.

Hint: Enter ‘1’ when you always want each row’s output to by printed on two lines. Enter ‘100’ when you want each row printed on only one line. This may cause some columns to be miss-aligned.

Show Notes

Check to display the notes at the end of each report.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Report Options – Decimal Places

Criterion - Z-Value

Specifies the number of decimal places used.

Report Options – Page Title

Report Page Title

Specify a page title to be displayed in report headings.

ROC Curve Tab

The options on this panel control the appearance of the ROC Curve.

Vertical and Horizontal Axis

Label

This is the text of the label. Press the button on the right of the field to specify the font of the text.

Tick Label Settings

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the tick labels along each axis.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

ROC Curve Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the ROC style file is used. These style files are created in the Scatter Plot procedure.

Calculation Points

The range of a criterion variable is divided into this many intervals and a two-by-two table is calculated. Note that each interval ends with a cutoff point.

ROC Curve Settings - Legend

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. Click the button on the right to specify the font size, color, and style of the legend text.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the response variable name. Press the button on the right of the field to specify the font of the text.

Lines for ROC Curves

These options set the color, width, and pattern of the up to fifteen lines representing the criterion variables. Note that the color of the 45 degree line is specified in the group immediately after the

criterion variables. For example, if you had three criterion variables, the color of the Group 4 option would be the color of the 45 degree line.

Double-clicking the line, or clicking the button to the right of the symbol, brings up a line specification window. This window lets you specify the characteristics of each line in detail.

Color

The color of the line.

Width

The width of the line.

Pattern

The line pattern (solid, dot, dash, etc.).

Storage Tab

Various proportions may be stored on the current database for further analysis. These options let you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database each time you run the procedure.

Note that the variables you specify must already have been named on the current database.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Criterion Values for Storage

Criterion Value List

Specify the specific values of the criterion variable to be used for storing the data back on the spreadsheet. Enter 'Data' when you want the values in the database to be used. You can enter a list using numbers separated by blanks such as '1 2 3 4 5' or you can use the 'xx TO yy BY inc' syntax or you can use the 'xx:yy(inc)' syntax.

For example, entering '1 TO 10 BY 3' or '1:10(3)' is the same as entering '1 4 7 10'.

Storage Variables

Store Variable Name in to Store NPV 2 in

Specify the variable to receive these values.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – ROC Curve for a Paired Design

This section presents an example of how to generate an ROC curve for the RMSF data contained in the ROC database. This is an example of data from a paired designed.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the ROC Curves window.

1 Open the ROC dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **ROC.s0**.
- Click **Open**.

2 Open the ROC Curves window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **ROC Curves**. The ROC Curves procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the ROC Curves window, select the **Variables tab**.
- Set the **Actual Condition Variable** to **Fever**.
- Set the **Positive Condition Value** to **1**.
- Set the **Actual Condition Prevalence** to **0.10**.
- Set the **Cost Benefit Ratios** to **1.1 1.3 1.5 1.7**.
- Set the **Criterion Variable(s)** to **Sodium1, Sodium2**.
- Set the **Test Direction** to **Low X = Positive**.
- Set the **Criterion List** to **120 to 140 by 5**.
- Set the **Max Equivalence Difference** to **0.05**.

4 Specify the reports.

- On the ROC Curves window, select the **Reports tab**.
- Check all reports and plots.
- Set the **Skip Line After** option to **20**.

5 Specify the Lines.

- On the ROC Curves window, select the **ROC Curve tab**.
- Click on the **Line 1** arrow. Set the **width** to **60**.
- Click on the **Line 2** arrow. Set the **width** to **60**.
- Click on the **Line 3** arrow. Set the **width** to **60**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

ROC Data using the Empirical ROC Curve

ROC Data for Condition = Fever using the Empirical ROC Curve

Sodium1 Cutoff Value	Count +|P A	Count +|A B	Count -|P C	Count -|A D	Sensitivity A/(A+C)	C/(A+C)	False+ B/(B+D)	Specificity D/(B+D)
120.00	0	0	21	24	0.00000	1.00000	0.00000	1.00000
125.00	2	0	19	24	0.09524	0.90476	0.00000	1.00000
130.00	11	1	10	23	0.52381	0.47619	0.04167	0.95833
135.00	18	6	3	18	0.85714	0.14286	0.25000	0.75000
140.00	21	19	0	5	1.00000	0.00000	0.79167	0.20833

ROC Data for Condition = Fever using the Empirical ROC Curve

Sodium2 Cutoff Value	Count +|P A	Count +|A B	Count -|P C	Count -|A D	Sensitivity A/(A+C)	C/(A+C)	False+ B/(B+D)	Specificity D/(B+D)
120.00	0	0	21	24	0.00000	1.00000	0.00000	1.00000
125.00	4	1	17	23	0.19048	0.80952	0.04167	0.95833
130.00	12	3	9	21	0.57143	0.42857	0.12500	0.87500
135.00	18	10	3	14	0.85714	0.14286	0.41667	0.58333
140.00	21	17	0	7	1.00000	0.00000	0.70833	0.29167

Notes:
A is the number of subjects with a POSITIVE test when the condition was PRESENT.
B is the number of subjects with a POSITIVE test when the condition was ABSENT.
C is the number of subjects with a NEGATIVE test when the condition was PRESENT.
D is the number of subjects with a NEGATIVE test when the condition was ABSENT.
Sensitivity is the Pr(Positive Test|Condition Present).
False+ is the Pr(Positive Test|Condition Absent).
Specificity is the Pr(Negative Test|Condition Absent).

The report displays the numeric information used to generate the empirical ROC curve.

Cutoff

The cutoff values of the criterion variable as set in the Criterion List option of the Reports panel.

A B C D

These four columns give the counts of the two-by-two tables that are formed at each of the corresponding cutoff points.

Sensitivity A/(A+C)

This is the proportion of those that had the disease that were correctly diagnosed by the test.

C/(A+C)

This is the proportion of those that had the disease that were incorrectly diagnosed.

False + B/(B+D)

The proportion of those who did not have the disease who were incorrectly diagnosed by the test as having it.

Specificity D/(B+D)

This is the proportion of those who did not have the disease who were correctly diagnosed as such.

ROC Data using the Binormal ROC Curve

ROC Data for Condition = Fever using the Binormal ROC Curve

Sodium1 Cutoff Value	Count +|P A	Count +|A B	Count -|P C	Count -|A D	Sensitivity A/(A+C)	C/(A+C)	False+ B/(B+D)	Specificity D/(B+D)
120.00	0	0	21	24	0.00751	1.00000	0.00000	1.00000
125.00	2	0	19	24	0.09734	0.90476	0.00000	0.99957
130.00	11	1	10	23	0.43561	0.47619	0.04167	0.97664
135.00	18	6	3	18	0.83464	0.14286	0.25000	0.74153
140.00	21	19	0	5	0.98246	0.00000	0.79167	0.24423

ROC Data for Condition = Fever using the Binormal ROC Curve

Sodium2 Cutoff Value	Count +|P A	Count +|A B	Count -|P C	Count -|A D	Sensitivity A/(A+C)	C/(A+C)	False+ B/(B+D)	Specificity D/(B+D)
120.00	0	0	21	24	0.01869	1.00000	0.00000	0.99949
125.00	4	1	17	23	0.14344	0.80952	0.04167	0.98899
130.00	12	3	9	21	0.48070	0.42857	0.12500	0.90223
135.00	18	10	3	14	0.83352	0.14286	0.41667	0.61742
140.00	21	17	0	7	0.97642	0.00000	0.70833	0.24291

The report displays the numeric information used to generate the binormal ROC curve.

Cutoff

The cutoff values of the criterion variable as set in the Criterion List option of the Reports panel.

A B C D

These four columns give the counts of the two-by-two tables that are formed at each of the corresponding cutoff points.

Sensitivity A/(A+C)

This is the proportion of those that had the disease that were correctly diagnosed by the test. Note that these values are based on the binormal model.

C/(A+C)

This is the proportion of those that had the disease that were incorrectly diagnosed. Note that these values are based on the binormal model.

False + B/(B+D)

The proportion of those who did not have the disease who were incorrectly diagnosed by the test as having it. Note that these values are based on the binormal model.

Specificity D/(B+D)

This is the proportion of those who did not have the disease who were correctly diagnosed as such. Note that these values are based on the binormal model.

Cost-Benefit Analysis – Empirical Curve

Cost - Benefit Analysis for Condition = Fever with Prevalence = 0.1 using Empirical Curve

Sodium1 Cutoff Value	Sensitivity	Specificity	Cost - Benefit When Ratio = 1.1000	Cost - Benefit When Ratio = 1.3000	Cost - Benefit When Ratio = 1.5000	Cost - Benefit When Ratio = 1.7000
120.00	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
125.00	0.0952	1.0000	0.0952	0.0952	0.0952	0.0952
130.00	0.5238	0.9583	0.1113	0.0363	-0.0387	-0.1137
135.00	0.8571	0.7500	-1.6179	-2.0679	-2.5179	-2.9679
140.00	1.0000	0.2083	-6.8375	-8.2625	-9.6875	-11.1125

Cost - Benefit Analysis for Condition = Fever with Prevalence = 0.1 using Empirical Curve

Sodium2 Cutoff Value	Sensitivity	Specificity	Cost - Benefit When Ratio = 1.1000	Cost - Benefit When Ratio = 1.3000	Cost - Benefit When Ratio = 1.5000	Cost - Benefit When Ratio = 1.7000
120.00	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
125.00	0.1905	0.9583	-0.2220	-0.2970	-0.3720	-0.4470
130.00	0.5714	0.8750	-0.6661	-0.8911	-1.1161	-1.3411
135.00	0.8571	0.5833	-3.2679	-4.0179	-4.7679	-5.5179
140.00	1.0000	0.2917	-6.0125	-7.2875	-8.5625	-9.8375

Notes:
The cost-benefit ratio is the ratio of the net cost when the condition is absent to the net cost when it is present.
Select the cutoff value for which the computed cost value is maximized (or minimized).
Prevalence is the actual probability of the condition in the population.

The report displays the numeric information used to generate the ROC curve.

Cutoff

The cutoff values of the criterion variable as set in the Criterion List option on the Reports tab.

Sensitivity

This is the proportion of those that had the disease that were correctly diagnosed by the test.

Specificity

This is the proportion of those who did not have the disease who were correctly diagnosed as such.

Cost-Benefit When Ratio = 1.1

The cost-benefit ratio is the ratio of the net cost when the condition is absent to the net cost when it is present. The optimum cutoff value is that one at which the computed cost value is maximized (or minimized).

Cost-Benefit Analysis – Binormal Curve

Cost - Benefit Analysis for Condition = Fever with Prevalence = 0.1 using Binormal Curve

Sodium1 Cutoff Value	Sensitivity	Specificity	Cost - Benefit When Ratio = 1.1000	Cost - Benefit When Ratio = 1.3000	Cost - Benefit When Ratio = 1.5000	Cost - Benefit When Ratio = 1.7000
120.00	0.0075	1.0000	0.0075	0.0075	0.0075	0.0075
125.00	0.0973	0.9996	0.0930	0.0922	0.0915	0.0907
130.00	0.4356	0.9766	0.2044	0.1623	0.1203	0.0783
135.00	0.8346	0.7415	-1.7242	-2.1895	-2.6547	-3.1200
140.00	0.9825	0.2442	-6.4997	-7.8600	-9.2204	-10.5808

Cost - Benefit Analysis for Condition = Fever with Prevalence = 0.1 using Binormal Curve

Sodium2 Cutoff Value	Sensitivity	Specificity	Cost - Benefit When Ratio = 1.1000	Cost - Benefit When Ratio = 1.3000	Cost - Benefit When Ratio = 1.5000	Cost - Benefit When Ratio = 1.7000
120.00	0.0187	0.9995	0.0137	0.0127	0.0118	0.0109
125.00	0.1434	0.9890	0.0345	0.0146	-0.0052	-0.0250
130.00	0.4807	0.9022	-0.4872	-0.6632	-0.8392	-1.0151
135.00	0.8335	0.6174	-2.9540	-3.6427	-4.3313	-5.0200
140.00	0.9764	0.2429	-6.5188	-7.8816	-9.2443	-10.6071

The report displays the numeric information used to generate the ROC curve.

Cutoff

The cutoff values of the criterion variable as set in the Criterion List option on the Reports tab.

Sensitivity

This is the proportion of those that had the disease that were correctly diagnosed by the test.

Specificity

This is the proportion of those who did not have the disease who were correctly diagnosed as such.

Cost-Benefit When Ratio = 1.1

The cost-benefit ratio is the ratio of the net cost when the condition is absent to the net cost when it is present. The optimum cutoff value is that one at which the computed cost value is maximized (or minimized).

Predicted Value Section – Empirical Method

Predictive Value Section for Fever using the Empirical ROC Curve

Sodium1 Cutoff Value	Sensitivity	Specificity	Likelihood Ratio	Prev. PPV	= 0.47 NPV	Prev. PPV	= 0.10 NPV
120.00	0.00000	1.00000		0.00000	0.53333	0.00000	0.90000
125.00	0.09524	1.00000		1.00000	0.55814	1.00000	0.90865
130.00	0.52381	0.95833	12.57143	0.91667	0.69697	0.58278	0.94768
135.00	0.85714	0.75000	3.42857	0.75000	0.85714	0.27586	0.97927
140.00	1.00000	0.20833	1.26316	0.52500	1.00000	0.12308	1.00000

Predictive Value Section for Fever using the Empirical ROC Curve

Sodium2 Cutoff Value	Sensitivity	Specificity	Likelihood Ratio	Prev. PPV	= 0.47 NPV	Prev. PPV	= 0.10 NPV
120.00	0.00000	1.00000		0.00000	0.53333	0.00000	0.90000
125.00	0.19048	0.95833	4.57143	0.80000	0.57500	0.33684	0.91420
130.00	0.57143	0.87500	4.57143	0.80000	0.70000	0.33684	0.94839
135.00	0.85714	0.58333	2.05714	0.64286	0.82353	0.18605	0.97351
140.00	1.00000	0.29167	1.41176	0.55263	1.00000	0.13559	1.00000

Notes:
Sensitivity is the Pr(Positive Test|Condition Present).
Specificity is the Pr(Negative Test|Condition Absent).
Likelihood Ratio is the ratio Pr(Positive Test|Condition Present)/Pr(Positive Test|Condition Absent).
Prev stands for the prevalence of the disease. The first value is from the data. The other was input.
PPV or Positive Predictive Value is the Pr(Condition Present|Positive Test).
NPV or Negative Predictive Value is the Pr(Condition Absent|Negative Test).

The report displays the information to assess the predicted value of the diagnostic test.

Cutoff

The cutoff values of the criterion variable as set in the Criterion List option on the Reports tab.

Sensitivity

This is the proportion of those that had the disease that were correctly diagnosed by the test.

Specificity

This is the proportion of those who did not have the disease who were correctly diagnosed as such.

Likelihood Ratio

The likelihood ratio statistic measures the value of the test for increasing certainty about a positive diagnosis. It is calculated as follows:

$$LR = \frac{\Pr(\textit{positive test}|\textit{disease})}{\Pr(\textit{positive test}|\textit{no disease})} = \frac{\textit{sensitivity}}{1-\textit{specificity}}$$

Prev = x.xxxx PPV

The values of PPV for the two prevalence values. The first prevalence value is the one that was calculated from the data. The second prevalence value was set by the user.

Prev = x.xxxx NPV

The values of NPV for the two prevalence values. The first prevalence value is the one that was calculated from the data. The second prevalence value was set by the user.

Predicted Value Section – Binormal Method

Predictive Value Section for Fever using the Empirical ROC Curve

Sodium1 Cutoff Value	Sensitivity	Specificity	Likelihood Ratio	Prev. PPV	= 0.47 NPV	Prev. PPV	= 0.10 NPV
120.00	0.00751	1.00000	5003.14160	0.99977	0.53521	0.99820	0.90068
125.00	0.09734	0.99957	223.92737	0.99492	0.55860	0.96136	0.90881
130.00	0.43561	0.97664	18.65033	0.94226	0.66416	0.67451	0.93966
135.00	0.83464	0.74153	3.22914	0.73860	0.83673	0.26405	0.97582
140.00	0.98246	0.24423	1.29995	0.53215	0.94088	0.12621	0.99208

Predictive Value Section for Fever using the Empirical ROC Curve

Sodium2 Cutoff Value	Sensitivity	Specificity	Likelihood Ratio	Prev. PPV	= 0.47 NPV	Prev. PPV	= 0.10 NPV
120.00	0.01869	0.99949	36.74964	0.96984	0.53790	0.80328	0.90164
125.00	0.14344	0.98899	13.02933	0.91936	0.56888	0.59145	0.91221
130.00	0.48070	0.90223	4.91677	0.81140	0.66506	0.35330	0.93989
135.00	0.83352	0.61742	2.17868	0.65592	0.80911	0.19490	0.97091
140.00	0.97642	0.24291	1.28969	0.53018	0.92170	0.12534	0.98933

The report displays the information to assess the predicted value of the diagnostic test.

Cutoff

The cutoff values of the criterion variable as set in the Criterion List option on the Reports tab.

Sensitivity

This is the proportion of those that had the disease that were correctly diagnosed by the test.

Specificity

This is the proportion of those who did not have the disease who were correctly diagnosed as such.

Likelihood Ratio

The likelihood ratio statistic measures the value of the test for increasing certainty about a positive diagnosis. It is calculated as follows:

$$LR = \frac{\Pr(\textit{positive test}|\textit{disease})}{\Pr(\textit{positive test}|\textit{no disease})} = \frac{\textit{sensitivity}}{1 - \textit{specificity}}$$

Prev = x.xxxx PPV

The values of PPV for the two prevalence values. The first prevalence value is the one that was calculated from the data. The second prevalence value was set by the user.

Prev = x.xxxx NPV

The values of NPV for the two prevalence values. The first prevalence value is the one that was calculated from the data. The second prevalence value was set by the user.

Area Under Curve Hypothesis Tests

Empirical Area Under Curve Analysis for Condition = Fever

Criterion	Empirical Estimate of AUC	AUC's Standard Error	Z-Value to Test AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalence of Fever	Count
Sodium1	0.87500	0.05052	7.42	0.0000	0.0000	0.46667	45
Sodium2	0.80754	0.06431	4.78	0.0000	0.0000	0.46667	45

Notes:
1. This approach underestimates AUC when there are only a few (3 to 7) unique criterion values.
2. The AUCs, SEs, and hypothesis tests above use the empirical approach for correlated (paired) samples developed by DeLong, DeLong, and Clarke-Pearson.
3. The Z-Value compares the AUC to 0.5, since the AUC of a 'useless' criterion is 0.5. The one-sided test is usually used here since your only interest is that the criterion is better than 'useless'.
4. The Z test used here is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.

Binormal Area Under Curve Analysis for Condition = Fever

Criterion	Binormal Estimate of AUC	AUC's Standard Error	Z-Value to Test AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalence of Fever	Count
Sodium1	0.87720	0.03995	4.70	0.0000	0.0000	0.46667	45
Sodium2	0.81350	0.06379	3.12	0.0009	0.0018	0.46667	45

Notes:
1. The AUCs, SEs, and hypothesis tests above use the binormal approach given by McClish (1989).
2. The Z-Value compares the AUC to 0.5, since the AUC of a 'useless' criterion is 0.5. The one-sided test is usually used here since your only interest is that the criterion is better than 'useless'.
3. The Z tests used here are only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.
4. The Z tests use a logistic-type transformation to achieve better normality.

These reports display areas under the ROC curve and associated standard errors and hypotheses tests for each of the criterion variables. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method.

The one-sided and two-sided hypothesis tests test the hypothesis that the diagnostic test is better than flipping a coin to make the diagnosis. The actual formulas used where presented earlier in this chapter.

Area Under Curve Confidence Intervals

Empirical Confidence Interval of AUC for Condition = Fever

Criterion	Empirical Estimate of AUC	AUC's Standard Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit	Prevalence of Fever	Count
Sodium1	0.87500	0.05052	0.73131	0.94432	0.46667	45
Sodium2	0.80754	0.06431	0.63966	0.90188	0.46667	45

Notes:
1. This approach underestimates AUC when there are only a few (3 to 7) unique criterion values.
2. The AUCs, SEs, and hypothesis tests above use the nonparametric approach developed by DeLong, DeLong, and Clarke-Pearson.
3. The confidence interval is based on the transformed AUC as given by Zhou et al (2002).
4. This method is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.

Binormal Confidence Interval of AUC for Condition = Fever

Criterion	Binormal Estimate of AUC	AUC's Standard Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit	Prevalence of Fever	Count
Sodium1	0.87720	0.03995	0.77143	0.93580	0.46667	45
Sodium2	0.81350	0.06379	0.64553	0.90640	0.46667	45

Notes:
1. The AUCs, SEs, and hypothesis tests above use the binormal approach given by McClish (1989).
2. The confidence intervals given here are only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.
3. The confidence interval use a logistic-type transformation to achieve better normality.

These reports display areas under the ROC curve and associated standard errors and confidence intervals for each of the criterion variables. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method. The actual formulas used where presented earlier in this chapter.

Tests of (AUC1-AUC2) = 0

Empirical Test of (AUC1 - AUC2) = 0 for Condition = Fever

Criterions 1,2	AUC1	AUC2	Difference Value	Difference Std Error	Difference Percent	Z-Value	Prob Level
Sodium1, Sodium2	0.87500	0.80754	0.06746	0.02130	-7.71	3.17	0.0015
Sodium2, Sodium1	0.80754	0.87500	-0.06746	0.02130	8.35	-3.17	0.0015

Notes:
1. This approach underestimates AUC when there are only a few (3 to 7) unique criterion values.
2. The AUCs, SEs, and hypothesis tests above use the nonparametric approach developed by DeLong, DeLong, and Clarke-Pearson.
3. This method is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.

Binormal Test of (AUC1 - AUC2) = 0 for Condition = Fever

Criterions 1,2	AUC1	AUC2	Difference Value	Difference Std Error	Difference Percent	Z-Value	Prob Level
Sodium1, Sodium2	0.87720	0.81350	0.06371	0.02717	-7.26	4.61	0.0000
Sodium2, Sodium1	0.81350	0.87720	-0.06371	0.02717	7.83	-4.61	0.0000

Notes:
1. The AUCs, SEs, and hypothesis tests above use the binormal approach.
2. The z-test is based on a logistic-type transformation of the areas.

These reports display the results of hypothesis tests concerning the equality of two AUCs. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method. The actual formulas used where presented earlier in this chapter.

Variances and Covariances of (AUC1-AUC2) = 0

Empirical Test Variances and Covariances for Condition = Fever

Criterions 1,2	AUC1	AUC2	AUC1 Variance	AUC2 Variance	AUC1,AUC2 Covariance	Difference Variance
Sodium1, Sodium2	0.87500	0.80754	0.00255	0.00414	0.00312	0.00045
Sodium2, Sodium1	0.80754	0.87500	0.00414	0.00255	0.00312	0.00045

Notes:
1. These AUCs, variances, and covariances use the nonparametric approach developed by DeLong, DeLong, and Clarke-Pearson.
2. This method is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.

Binormal Test Variances and Covariances for Condition = Fever

Criterions 1,2	AUC1	AUC2	AUC1 Variance	AUC2 Variance	AUC1,AUC2 Covariance	Difference Variance
Sodium1, Sodium2	0.87720	0.81350	0.00160	0.00407	0.00246	0.00074
Sodium2, Sodium1	0.81350	0.87720	0.00407	0.00160	0.00246	0.00074

Notes:
1. The AUCs, SEs, and hypothesis tests above use the binormal approach.
2. The z-test is based on a logistic-type transformation of the areas.

These reports display the variances and covariances associated with the hypothesis tests given in the last set of reports. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method. The actual formulas used where presented earlier in this chapter.

Confidence Intervals of (AUC1-AUC2) = 0

Empirical Confidence Intervals of Differences in AUCs for Condition = Fever

Criterions 1,2	AUC1	AUC2	Difference Value	Difference Std Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
Sodium1, Sodium2	0.87500	0.80754	0.06746	0.02130	0.02571	0.10921
Sodium2, Sodium1	0.80754	0.87500	-0.06746	0.02130	-0.10921	-0.02571

Notes:
1. This approach underestimates AUC when there are only a few (3 to 7) unique criterion values.
2. The AUCs, SEs, and confidence limits above use the nonparametric approach developed by DeLong, DeLong, and Clarke-Pearson.
3. This method is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.

Binormal Confidence Interval of Difference in AUCs for Condition = Fever

Criterions 1,2	AUC1	AUC2	Difference Value	Difference Std Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit
Sodium1, Sodium2	0.87720	0.81350	0.06371	0.02717	0.01046	0.11696
Sodium2, Sodium1	0.81350	0.87720	-0.06371	0.02717	-0.11696	-0.01046

Notes:
1. The AUCs, SEs, and hypothesis tests above use the binormal approach.
2. The z-test is based on a logistic-type transformation of the areas.

These reports display the confidence intervals for difference between a pair of AUCs. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method. The actual formulas used where presented earlier in this chapter.

Equivalence Tests Comparing AUC1 and AUC2

Empirical Equivalence Test of the AUCs for Condition = Fever

Criterion Variable1, Variable2	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Sodium1, Sodium2	0.7938	0.03242	0.10250	-0.05000	0.05000	Cannot reject H0
Sodium2, Sodium1	0.7938	-0.10250	-0.03242	-0.05000	0.05000	Cannot reject H0

Notes:
1. This approach underestimates AUC when there are only a few (3 to 7) unique criterion values.
2. The AUCs, SEs, and confidence limits above use the empirical approach developed by DeLong, DeLong, and Clarke-Pearson.
3. This method is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.
4. Equivalence means that Test2 does not differ from Test1 except for a small, negligible amount which we call the 'Equivalence Bound'.

Binormal Equivalence Test of the AUCs for Condition = Fever

Criterion Variable1, Variable2	Prob Level	Lower 90.0% Conf. Limit	Upper 90.0% Conf. Limit	Lower Equiv. Bound	Upper Equiv. Bound	Reject H0 and Conclude Equivalence at the 5.0% Significance Level
Sodium1, Sodium2	0.0000	0.01902	0.10839	-0.05000	0.05000	Yes: (AUC1-AUC2)<0.05
Sodium2, Sodium1	0.6930	-0.10839	-0.01902	-0.05000	0.05000	Cannot reject H0

Notes:
1. The AUCs, SEs, and confidence limits above use the binormal approach.
2. Equivalence means that Test2 does not differ from Test1 except for a small, negligible amount which we call the 'Equivalence Bound'.
3. The logistic-type transformation is not used in these calculations.

These reports display the results of an equivalence test. This hypothesis test tests whether the two diagnostics are equivalence in the sense that their AUC's are no more different than the maximum amount specified. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method.

Often, you want to show that one diagnostic test is equivalent to another. In this case, you would use a test of equivalence.

Noninferiority Tests Comparing AUC1 and AUC2

Empirical Noninferiority Test of the AUCs for Condition = Fever

Criterion Variable1, Variable2	Prob Level	1-Sided 95.0% Conf. Limit	Noninferiority Bound	Reject H0 and Conclude Noninferiority at the 5.0% Significance Level
Sodium1, Sodium2	0.0000	0.03242	0.05000	Yes: (AUC1-AUC2)<0.05
Sodium2, Sodium1	0.7938	-0.10250	0.05000	Cannot reject H0

Notes:
1. This approach underestimates AUC when there are only a few (3 to 7) unique criterion values.
2. The AUCs, SEs, and confidence limits above use the empirical approach developed by DeLong, DeLong, and Clarke-Pearson.
3. This method is only accurate for larger samples with at least 30 subjects with, and another 30 subjects without, the condition of interest.
4. Noninferiority means that Test2 is no worse than Test1 except for a small, negligible amount which we call the 'Noninferiority Bound'.

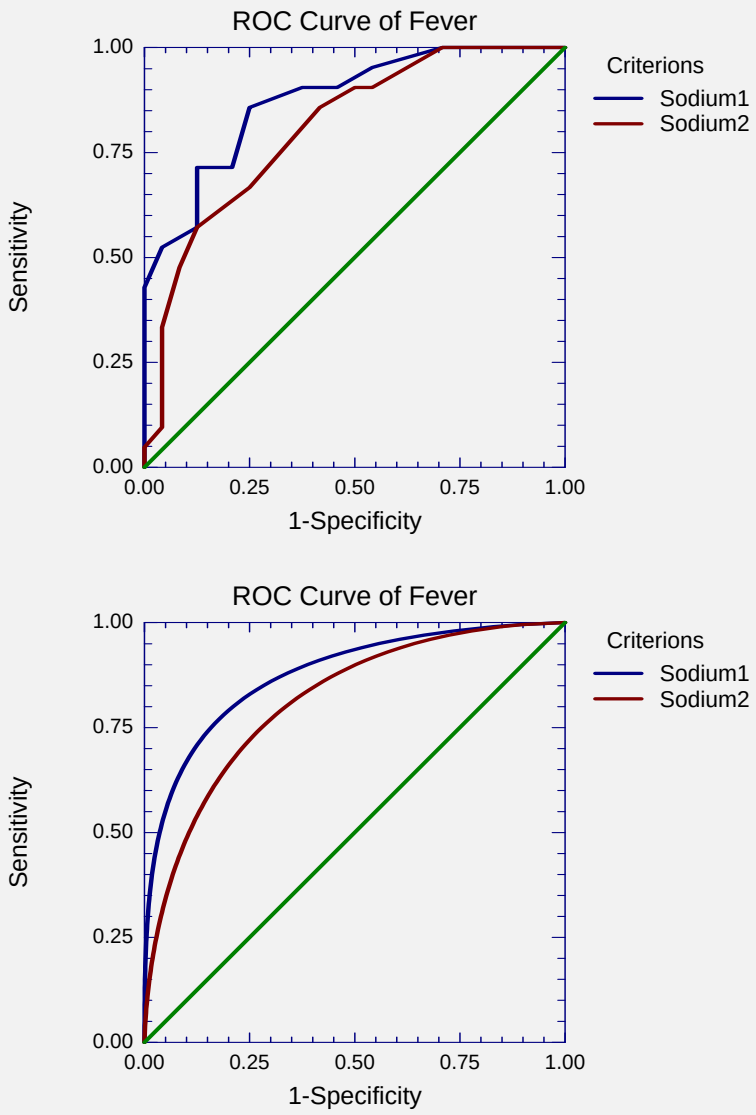
Binormal Noninferiority Test of the AUCs for Condition = Fever

Criterion Variable1, Variable2	Prob Level	1-Sided 95.0% Conf. Limit	Noninferiority Bound	Reject H0 and Conclude Noninferiority at the 5.0% Significance Level
Sodium1, Sodium2	0.0000	0.01902	0.05000	Yes: (AUC1-AUC2)<0.05
Sodium2, Sodium1	0.6930	-0.10839	0.05000	Cannot reject H0

Notes:
1. The AUCs, SEs, and confidence limits above use the binormal approach.
2. Noninferiority means that Test2 is no worse than Test1 except for a small, negligible amount which we call the 'Noninferiority Bound'.
3. The logistic-type transformation is not used in these calculations.

These reports display the results of a noninferiority test. This hypothesis test tests whether diagnostic test 2 is no worse than diagnostic test 1 in the sense that AUC2 is not less than AUC1 by more than a small amount. The first report is based on the empirical ROC method and the second report is based on the binormal ROC method.

ROC Plot Section



Both the empirical and binormal ROC curves are displayed here. The empirical curve is shown first. It always has a 'zig-zag' pattern. The smooth, binormal ROC curve is shown second.

The ROC curves plot the proportion of those who actually had the disease who were correctly diagnosed on the vertical axis versus the proportion of those who did not have the disease who were falsely diagnosed as having it on the horizontal axis. Hence, an optimum test procedure is one whose ROC curve proceeds from the lower-left corner vertically until it reaches the top and then horizontally across the top to the right side. The 45 degree line represents what you would expect from a chance (flip of the coin) classification procedure.

When you are comparing two curves as in this example, you would generally take the outside curve (the one furthest from the middle line). However, it is possible for the curves to cross so that one test is optimum in a certain range but not in another.

Example 2 – Validation Using Zhou et al. (2002)

Zhou et al. (2002) page 175 presents an example comparing the results of two mammography tests: plain film and digitized film. In this example, both tests were administered to 58 people, yielding a paired design. The results given in Zhou (2002) contain a few typos. We have obtained the corrected results from the authors, which are as follows:

Test	AUC	SE
Plain	0.83504	0.06581
Digitized	0.84701	0.05987

The data for this example is contained in the dataset ZHOU 175.S0 . You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the ROC Curves window.

1 Open the ZHOU 175 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the Data subdirectory of your NCSS directory.
- Click on the file **ZHOU 175.S0**.
- Click **Open**.

2 Open the ROC Curves window.

- On the menus, select **Analysis**, then **Diagnostic Tests**, then **ROC Curves**. The ROC Curves procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the ROC Curves window, select the **Variables tab**.
- Set the **Actual Condition Variable** to **CANCER**.
- Set the **Positive Condition Value** to **1**.
- Set the **Actual Condition Prevalence** to **0.10**.
- Set the **Cost Benefit Ratios** to **1**.
- Set the **Criterion Variable(s)** to **Plainfilm-Digifilm**.
- Set the **Test Direction** to **High X = Positive**.
- Set the **Criterion List** to **1 2 3 4**.
- Set the **Frequency Variable** to **Count**.
- Set the **Max Equivalence Difference** to **0.05**.

4 Specify the reports.

- On the ROC Curves window, select the **Reports tab**.
- Check the box next to **Area Under Curve (AUC): One Curve**.
- Set the **Skip Line After** option to **20**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Validation Output

Empirical Area Under Curve Analysis for Condition = Cancer

Criterion	Empirical Estimate of AUC	AUC's Standard Error	Z-Value to Test AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalence of Fever	Count
PlainFilm	0.83504	0.06581	5.09	0.0000	0.0000	0.22414	58
DigiFilm	0.84701	0.05987	5.80	0.0000	0.0000	0.22414	58

Note that the estimated AUC’s and standard errors match those given by Zhou (2002) exactly.

Chapter 550

Distribution (Weibull) Fitting

Introduction

This procedure estimates the parameters of the exponential, extreme value, logistic, log-logistic, lognormal, normal, and Weibull probability distributions by maximum likelihood. It can fit complete, right censored, left censored, interval censored (readout), and grouped data values. It also computes the nonparametric Kaplan-Meier and Nelson-Aalen estimates of survival and associated hazard rates. It outputs various statistics and graphs that are useful in reliability and survival analysis. When the choice of the probability distribution is in doubt, the procedure helps select an appropriate probability distribution from those available.

Features of this procedure include:

1. Probability plotting, hazard plotting, and reliability plotting for the common life distributions. The data may be any combination of complete, right censored, left censored, and interval censored data.
2. Maximum likelihood and probability plot estimates of distribution parameters, percentiles, reliability (survival) functions, hazard rates, and hazard functions.
3. Confidence intervals for distribution parameters and percentiles.
4. Nonparametric estimates of survival using the Kaplan-Meier procedure.

Overview of Survival and Reliability Analysis

This procedure may be used to conduct either survival analysis or reliability analysis. When a study concerns a biological event associated with the study of animals (including humans), it is usually called *survival analysis*. When a study concerns machines in an industrial setting, it is usually called *reliability analysis*. Survival analysis emphasizes a nonparametric estimation approach (Kaplan-Meier estimation), while reliability analysis emphasizes a parametric approach (Weibull or lognormal estimation). In the rest of this chapter, when we refer to survival analysis, you can freely substitute ‘reliability’ for ‘survival.’ The two terms refer to the same type of analysis.

We will give a brief introduction to the subject in this section. For a complete account of survival analysis, we suggest the book by Klein and Moeschberger (1997).

Survival analysis is the study of the distribution of life times. That is, it is the study of the elapsed time between an initiating event (birth, start of treatment, diagnosis, or start of operation) and a

terminal event (death, relapse, cure, or machine failure). The data values are a mixture of complete (terminal event occurred) and censored (terminal event has not occurred) observations. From the data values, the survival analyst makes statements about the survival distribution of the failure times. This distribution allows questions about such quantities as survivability, expected life time, and mean time to failure to be answered.

Let *T* be the elapsed time until the occurrence of a specified event. The event may be death, occurrence of a disease, disappearance of a disease, appearance of a tumor, etc. The probability distribution of *T* may be specified using one of the following basic functions. Once one of these functions has been specified, the others may be derived using the mathematical relationships presented.

1. Probability density function, *f(t)*. This is the probability that an event occurs at time *t*.
2. Cumulative distribution function, *F(t)*. This is the probability that an individual survives until time *t.*

$$F(t) = \int_0^t f(x)dx$$

3. Survival function, *S(t)* or Reliability function, *R(t).* This is the probability that an individual survives beyond time *t*. This is usually the first quantity that is studied. It may be estimated using the nonparametric Kaplan-Meier curve or one of the parametric distribution functions.

$$R(t) = S(t) = \int_t^\infty f(x)dx = 1 - F(t)$$

$$S(t) = \exp\left[-\int_0^t h(x)dx \right] = \exp\left[- H(t) \right]$$

4. Hazard rate, *h(t)*. This is the probability that an individual at time *t* experiences the event in the next instant. It is a fundamental quantity in survival analysis. It is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, and the inverse of Mill's ratio in economics. The empirical hazard rate may be used to identify the appropriate probability distribution of a particular mechanism, since each distribution has a different hazard rate function. Some distributions have a hazard rate that decreases with time, others have a hazard rate that increases with time, some are constant, and some exhibit all three behaviors at different points in time.

$$h(t) = \frac{f(t)}{S(t)}$$

5. Cumulative hazard function, *H(t)*. This is integral of *h(t)* from 0 to *t*.

$$H(t) = \int_0^t h(x)dx = -\ln\left[S(t) \right]$$

Nonparametric Estimators of Survival

There are two competing nonparametric estimators of the survival distribution, *S(t)*, available in this procedure. The first is the common Kaplan-Meier Product limit estimator. The second is the Nelson-Aalen estimator of the cumulative hazard function, *H(t)*.

Kaplan-Meier Product-Limit Estimator

The most common nonparametric estimator of the survival function is called the Kaplan-Meier product limit estimator. This estimator is defined as follows in the range of time values for which there are data.

$$\hat{S}(t) = \begin{cases} 1 & if\ t < t_1 \\ \prod_{t_1 \le t} \left[1 - \frac{d_i}{Y_i} \right] & if\ t_1 \le t \end{cases}$$

In the above equation, d_i represents the number of deaths at time t_i and Y_i represents the number of individuals who are at risk at time t_i.

The variance of *S(t)* is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i < t} \frac{d_i}{Y_i(Y_i - d_i)}$$

The product limit estimator may be used to estimate the cumulative hazard function *H(t)* using the relationship

$$\hat{H}(t) = -\log[\hat{S}(t)]$$

Linear (Greenwood) Confidence Limits

This estimator may be used to create confidence limits for *S(t)* using the formula

$$\hat{S}(t) \pm z_{1-\alpha/2}\sigma_S(t)\hat{S}(t)$$

where

$$\sigma_S^2(t) = \frac{\hat{V}\left[\hat{S}(t)\right]}{\hat{S}^2(t)}$$

and *z* is the appropriate value from the standard normal distribution. We call this the *Linear (Greenwood) confidence interval*.

Log Hazard Confidence Limits

Better confidence limits may be calculated using the logarithmic transformation of the hazard functions. These limits are

$$\hat{S}(t)^{1/\theta}, \hat{S}(t)^{\theta}$$

where

$$\theta = \exp\left\{ \frac{z_{1-\alpha/2}\sigma_S(t)}{\log[\hat{S}(t)]} \right\}$$

ArcSine-Square Root Hazard Confidence Limits

Another set of confidence limits using an improving transformation is given by the (intimidating) formula

$$\sin^2\left\{\max\left[0, \arcsin(\hat{S}(t)^{1/2} - 0.5z_{1-\alpha/2}\sigma_S(t)\left(\frac{\hat{S}(t)}{1-\hat{S}(t)}\right)^{1/2}\right]\right\}$$
$$\le S(t) \le$$
$$\sin^2\left\{\min\left[\frac{\pi}{2}, \arcsin(\hat{S}(t)^{1/2} + 0.5z_{1-\alpha/2}\sigma_S(t)\left(\frac{\hat{S}(t)}{1-\hat{S}(t)}\right)^{1/2}\right]\right\}$$

Nelson-Aalen Hazard Confidence Limits

An alternative estimator of *H(t)*, which has better small sample size properties is the Nelson-Aalen estimator given by

$$\tilde{H}(t) = \begin{cases} 0 & if\ \ t < t_1 \\ \sum_{t_1 \le t} \frac{d_i}{Y_i} & if\ \ t_1 \le t \end{cases}$$

The variance of this estimate is given by the formula

$$\sigma_H^2(t) = \sum_{t_i \le t} \frac{(Y_i - d_i)d_i}{(Y_i - 1)Y_i^2}$$

The 100(1-alpha)% confidence limits for *H(t)* are calculated using

$$\tilde{H}(t)\exp\left(\pm z_{1-\alpha/2}\sigma_H(t) / \tilde{H}(t)\right)$$

This hazard function may be used to generate the Nelson-Aalen estimator of *S(t)* using the formula

$$\tilde{S}(t) = e^{\tilde{H}(t)}$$

Using these formulas, a fourth set of confidence limits for *S(t)* may be calculated as

$$\exp\left\{\tilde{H}(t) \pm z_{1-\alpha/2}\sigma_H(t)\right\}$$

Parametric Survival Distributions

This section presents the parametric probability distributions that may be analyzed with this procedure.

Normal Distribution

The normal distribution is one of the most commonly used in statistics. However, it is used infrequently as a lifetime distribution because it allows negative values while lifetimes are always positive. It has been found that the logarithms of failure times may be fit by the normal

distribution. Hence the lognormal has become a popular distribution in reliability work, while the normal has been put on the sideline.

The normal distribution is indexed by a location (M) and a scale (S) parameter. A threshold parameter is meaningless in this case, since it is an adjustment to the location parameter. Using these symbols, the normal density function may be written as

$$f(t|M,S) = \frac{1}{S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-M}{S}\right)^2}, \quad -\infty < M < \infty, S > 0, -\infty < t < \infty$$

Location Parameter - M

The location parameter of the normal distribution is often called the mean.

Scale Parameter - S

The scale parameter of the normal distribution is usually called the standard deviation.

Lognormal Distribution

The normal distribution is one of the most commonly used in statistics. Although the normal distribution itself does not often work well with time-to-failure data, it has been found that the logarithm of failure time often does. Hence the lognormal has become a popular distribution in reliability work.

The lognormal distribution is indexed by a shape (S) , a scale (M), and a threshold (D) parameter. Using these symbols, the three parameter *lognormal* density function may be written as

$$f(t|M,S,D) = \frac{1}{(t-D)S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(t-D)-M}{S}\right)^2}, -\infty < M < \infty, S > 0, -\infty < D < \infty, t > D$$

It is often more convenient to work with logarithms to the base 10 (denoted by *log*) rather than logarithms to the base e (denoted by *ln*). The *lognormal10* density function is written as

$$f(t|M,S,D) = \frac{1}{\ln(10)(t-D)S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(t-D)-M}{S}\right)^2}, -\infty < M < \infty, S > 0, -\infty < D < \infty, t > D$$

Shape Parameter - S

The shape parameter of the lognormal distribution of *t* is the standard deviation in the normal distribution of *ln(t-D)* or *log(t-D)*. That is, the scale parameter of the normal distribution is the shape parameter of the lognormal distribution.

Scale Parameter - M

The scale parameter of the lognormal distribution *t* is the mean in the normal distribution of *ln(t-D)* or *log(t-D)*. That is, the location parameter of the normal distribution is the scale parameter of the lognormal distribution.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. When *D* is set to zero, we obtain the two parameter lognormal distribution. When a positive value is given to *D*, we are inferring that no failures can occur between zero and *D*.

Reliability Function

The reliability (or survivorship) function, *R(t)*, gives the probability of surviving beyond time *t*. For the Lognormal distribution, the reliability function is

$$R(t) = 1 - \Phi\left(\frac{\ln(t-D) - M}{S}\right)$$

where $\Phi(z)$ is the standard normal distribution function.

The conditional reliability function, *R(t,T)*, may also be of interest. This is the reliability of an item given that it has not failed by time *T*. The formula for the conditional reliability is

$$R(t) = \frac{R(T+t)}{R(T)}$$

Hazard Function

The hazard function represents the instantaneous failure rate. For this distribution, the hazard function is

$$h(t) = \frac{f(t)}{R(t)}$$

Weibull Distribution

The Weibull distribution is named for Professor Waloddi Weibull whose papers led to the wide use of the distribution. He demonstrated that the Weibull distribution fit many different datasets and gave good results, even for small samples. The Weibull distribution has found wide use in industrial fields where it is used to model time to failure data.

The three parameter Weibull distribution is indexed by a shape (B) , a scale (C), and a threshold (D) parameter. Using these symbols, the three parameter Weibull density function may be written as

$$f(t|B,C,D) = \frac{B}{C}\left(\frac{t-D}{C}\right)^{(B-1)} e^{-\left(\frac{t-D}{C}\right)^{B}}, \quad B>0, C>0, -\infty < D < \infty, t > D.$$

The symbol *t* represents the random variable (usually elapsed time). The threshold parameter *D* represents the minimum value of *t* that can occur. Setting the threshold to zero results in the common, two parameter Weibull distribution.

Shape Parameter - B

The shape (or power) parameter controls the overall shape of the density function. Typically, this value ranges between 0.5 and 8.0. The estimated standard errors and confidence limits displayed by the program are only valid when $B > 2.0$.

One of the reasons for the popularity of the Weibull distribution is that it includes other useful distributions as special cases or close approximations. For example, if

B = 1 The Weibull distribution is identical to the exponential distribution.

B = 2 The Weibull distribution is identical to the Rayleigh distribution.

B = 2.5 The Weibull distribution approximates the lognormal distribution.

B = 3.6 The Weibull distribution approximates the normal distribution.

Scale Parameter - C

The scale parameter only changes the scale of the density function along the time axis. Hence, a change in this parameter has the same effect on the distribution as a change in the scale of time—for example, from days to months or from hours to days. However, it does not change the actual shape of the distribution.

C is known as the *characteristic life.* No matter what the shape, 63.2% of the population fails by $t = C+D$.

Some authors use $1/C$ instead of *C* as the scale parameter. Although this is arbitrary, we prefer dividing by the scale parameter since that is how you usually scale a set of numbers. For example, remember how you create a z-score when dealing with the normal data or create a percentage by dividing by the maximum.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. Often, this parameter is referred to as the *location* parameter. We use ‘threshold’ rather than ‘location’ to stress that this parameter sets the minimum time. We reserve ‘location’ to represent the center of the distribution. This is a fine point and we are not upset when people refer to this as the location parameter.

When *D* is set to zero, we obtain the two parameter Weibull distribution. It is possible, but unusual, for *D* to have a negative value. When using a search algorithm to find the estimated value of *D*, a nonzero value will almost certainly be found. However, you should decide physically if a zero probability of failure in the interval between 0 and *D* is truly warranted.

A downward or upward sloping tail on the Weibull probability plot or values of $B > 6.0$ are indications that a nonzero threshold parameter will produce a better fit to your data.

Negative values of *D* represent an amount of time that has been subtracted from the actual times. On the other hand, positive values of *D* represent a period of time between the starting point and when any failures can occur. For example, positive values of *D* may represent the amount of time between the production of an item and when it is placed in service.

Relationship to the Extreme Value Distribution

The extreme value distribution is directly related to the Weibull distribution. If $x = ln(t)$ and *t* follows the Weibull distribution, *x* follows the extreme value distribution.

Reliability Function

The reliability (or survivorship) function, *R(t)*, gives the probability of surviving beyond time *t*. For the Weibull distribution, the reliability function is

$$R(t) = e^{-\left(\frac{t-D}{C}\right)^B}$$

The reliability function is one minus the cumulative distribution function. That is,

$$R(t) = 1 - F(t)$$

The conditional reliability function, *R(t,T)*, may also be of interest. This is the reliability of an item given that it has not failed by time *T*. The formula for the conditional reliability is

$$R(t) = \frac{R(T+t)}{R(T)}$$

Hazard Function

The hazard function represents the instantaneous failure rate. For this distribution, the hazard function is

$$h(t) = \frac{f(t)}{R(t)} = \frac{B}{C}\left(\frac{t-D}{C}\right)^{B-1}$$

Depending on the values of the distribution's parameters, the Weibull's hazard function can be decreasing (when $B<1$), constant (when $B=1$ at $1/C$), or increasing (when $B>1$) over time.

Extreme Value Distribution

The extreme value distribution is occasionally used to model lifetime data. It is included here because of its close relationship to the Weibull distribution. It turns out that if *t* is distributed as a Weibull random variable, then *ln(t)* is distributed as the extreme value distribution.

The density of the extreme value distribution may be written as

$$f(t|M,S) = \frac{1}{S}\exp\left(\frac{t-M}{S}\right)\exp\left(-\exp\left(\frac{t-M}{S}\right)\right), \quad S>0$$

Exponential Distribution

The exponential distribution was one of the first distributions used to model lifetime data. It has now been superceded by the Weibull distribution, but is still used occasionally. The exponential distribution may be found from the Weibull distribution by setting $B = 1$.

The exponential distribution is a model for the life of products with a constant failure rate. The two parameter exponential distribution is indexed by both a scale and a threshold parameter. The density of the exponential distribution may be written as

$$f(t|S,D) = \frac{1}{S}\exp\left(-\frac{t-D}{S}\right), \quad S>0$$

Scale Parameter - S

The scale parameter changes the scale of the density function along the time axis. Hence, a change in this parameter has the same effect on the distribution as a change in the scale of time—for example, from days to months or from hours to days. However, it does not change the actual shape of the distribution.

Some authors use 1/*S* instead of *S* as the scale parameter. Although this is arbitrary, we prefer dividing by the scale parameter since that is how a set of numbers is usually scaled. For example, remember how z-scores are created when dealing with the normal distribution.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. Often, this parameter is referred to as the *location* parameter. We use 'threshold' rather than 'location' to stress that this parameter sets the minimum time. We reserve 'location' to represent the center of the distribution. This is a fine point and we are not upset when people refer to this as the location parameter.

When *D* is set to zero, we obtain the two parameter exponential distribution. It is possible, but unusual, for *D* to have a negative value. When using a search algorithm to find the estimated value of *D*, a nonzero value will almost certainly be found. However, you should decide physically if a zero probability of failure in the interval between 0 and *D* is truly warranted.

A downward or upward sloping tail on the exponential probability plot is an indication that a nonzero threshold parameter will produce a better fit to your data.

Negative values of *D* represent an amount of time that has been subtracted from the actual times. On the other hand, positive values of *D* represent a period of time between the starting point and when any failures can occur. For example, positive values of *D* may represent the amount of time between the production of an item and when it is placed in service.

Reliability Function

The reliability (or survivorship) function, *R(t)*, gives the probability of surviving beyond time *t*. For the exponential distribution, the reliability function is

$$R(t) = e^{-\left(\frac{t-D}{S}\right)}$$

Hazard Function

The hazard function represents the instantaneous failure rate. For this distribution, the hazard function is constant. The rate is given by the function

$$h(t) = \frac{f(t)}{R(t)} = \frac{1}{S}$$

Logistic Distribution

The density of the logistic distribution may be written as

$$f(t|M,S) = \frac{\exp\left(\frac{t-M}{S}\right)}{S\left[1+\exp\left(\frac{t-M}{S}\right)\right]^2}, \quad S>0$$

Log-logistic Distribution

The density of the log-logistic distribution may be written as

$$f(t|M,S,D) = \frac{\exp\left(\frac{\ln(t-D)-M}{S}\right)}{(t-D)S\left[1+\exp\left(\frac{\ln(t-D)-M}{S}\right)\right]^2}, \quad S>0$$

The log-logistic distribution is used occasionally to model lifetime data, but it is so similar to the lognormal and the Weibull distributions that it adds little and is thus often dropped from consideration.

Parameter Estimation

The parameters of the reliability distributions may be estimated by maximum likelihood or by applying least squares regression to the probability plot. The probability plot method is popular because it uses a nice graphic (the probability plot) which allows a visual analysis of the goodness of fit of the distribution to the data. Maximum likelihood estimation is usually favored by statisticians because it has been shown to be optimum in most situations and because it provides estimates of standard errors and confidence limits. However, there are situations in which maximum likelihood does not do as well as the regression approach. For example, maximum likelihood does not do a good job of estimating the threshold parameter. When you want to include the threshold parameter in your model, we suggest that you use the regression approach to estimate it and then treat the threshold value as a known quantity in the maximum likelihood estimation.

Maximum Likelihood

Maximum likelihood estimation consists of finding the values of the distribution parameters that maximize the log-likelihood of the data values. Loosely speaking, these are the values of the parameters which maximize the probability that the current set of data values occur.

The general form of the log-likelihood function is given by

$$L(\underline{P}) = \sum_F \ln\left(f(\underline{P},t_k)\right) + \sum_R \ln\left(S(\underline{P},t_k)\right) + \sum_L \ln\left(F(\underline{P},t_k)\right) + \sum_I \ln\left(f(\underline{P},t_{uk}) - f(\underline{P},t_{lk})\right)$$

where F stands for the set of failed items, R stands for the set of right censored items, L stands for the set of left censored items, and I stands for the set of interval censored items. In the case of

interval censored observations, t_{lk} represents the first time of the interval and t_{uk} represents the last time of the interval. Also, $\underline{P}$ represents one or two parameters as the case may be.

$L(\underline{P})$ is maximized using two numerical procedures. First, a recently developed method called *differential evolution* is used. This is a robust maximization procedure that only requires evaluations of the function, but not its derivatives. The solution of the differential evolution phase is then used as starting values for the Newton-Raphson algorithm. Newton-Raphson is used because it provides an estimate of the variance-covariance matrix of the parameters, which is needed in computing confidence limits. The amount of emphasis on the differential evolution phase as opposed to the Newton-Raphson phase may be controlled using the maximum number of iterations allowed for each. Numerical differentiation is used to compute the first and second order derivatives that are needed by the Newton-Raphson procedure.

Data Structure

Survival data is somewhat more difficult to enter because of the presence of various types of censoring.

Failed or Complete

A failed observation is one in which the time until the terminal event was measured exactly; for example, the machine stopped working or the mouse died of the disease being studied.

Right Censored

A right censored observation provides a lower bound for the actual failure time. All that is known is that the failure occurred (or will occur) at some point after the given time value. Right censored observations occur when a study is terminated before all items have failed. They also occur when an item fails due to an event other than the one of interest.

Left Censored

A left censored observation provides an upper bound for the actual failure time. All we know is that the failure occurred at some point before the time value. Left censoring occurs when the items are not checked for failure until some time after the study has begun. When a failed item is found, we do not know exactly when it failed, only that it was at some point before the left censor time.

Interval Censored or Readout

An interval censored observation is one in which we know that the failure occurred between two time values, but we do not know exactly when. This type of data is often called *readout* data. It occurs in situations where items are checked periodically for failures.

Sample Dataset

Most data sets require two (and often three) variables: the failure time variable, an optional censor variable indicating the type of censoring, and an optional count variable which gives the number of items occurring at that time. If the censor variable is omitted, all time values represent failed items. If the count variable is omitted, all counts are assumed to be one.

The table below shows the results of a study to test the failure rate of a particular machine. This particular experiment began with 30 items being tested. After the twelfth item failed at 152.7 hours, the experiment was stopped. The remaining eighteen observations were right censored. That is, we know that they will fail at some time in the future. These data are contained in the WEIBULL database.

WEIBULL dataset

Time	Censor	Count
12.5	1	1
24.4	1	1
58.2	1	1
68.0	1	1
69.1	1	1
95.5	1	1
96.6	1	1
97.0	1	1
114.2	1	1
123.2	1	1
125.6	1	1
152.7	1	1
152.7	0	18

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the probability distribution that is fit and the variables used in the analysis.

Time Variables

Time Variable

This variable contains the time values for both failed and censored observations. When interval (readout) data are used, this variable specifies the ending time of the interval.

Negative time values are treated as missing values. Zero time values are replaced by the value in the Zero field.

These time values represent elapsed times. If your dataset is made up of dates (such as the failure date), you must subtract the starting date from the failure date so that you can analyze the elapsed time.

Start Time Variable

This variable contains the starting time for interval (readout) data. Hence its value is only used when the row's censor value indicates an interval data type.

Negative time values are treated as missing values. Zero time values are replaced by the value in the Zero field.

Zero time Replacement Value

Under normal conditions, a respondent beginning the study is “alive” and cannot “die” until after some small period of time has elapsed. Hence, a time value of zero is not defined and is ignored (treated as a missing value). If a zero time value does occur in the database, it is replaced by this positive amount. If you do not want zero time values replaced, enter a “0.0” here.

This option is used when a “zero” on the database does not actually mean zero time. Instead, it means that the response occurred before the first reading was made and so the actual survival time is only known to be less.

Censor Variable

Censor Variable

The values in this optional variable indicate the type of censoring active for each row. Four possible data types may be entered: failed (complete), right censored, left censored, or interval. The values used to indicate each data type are specified in the four boxes to the right. These values may be text or numeric.

Failed

When this value is entered in the Censor Variable, the corresponding time value is treated as a failed observation. The value may be a number or a letter. We suggest the letter “F” when you are in doubt as to what to use.

A failed observation is one in which the time until the event of interest was measured exactly; for example, the machine stopped working or the mouse died of the disease being studied. The exact failure time is known.

Right

When this value is entered in the Censor Variable, the corresponding time value is treated as a right censored data value. The value may be a number or a letter. We suggest the letter “R” when you are in doubt as to what to use.

A right censored observation provides a lower bound for the actual failure time. All that is known is that the failure time occurred (or will occur) at some point after the given time value. Right censored observations occur when a study is terminated before all items have failed. They also occur when an item fails due to an event other than the one of interest.

Left

When this value is entered in the Censor Variable, the corresponding time value is treated as a left censored data value. The value may be a number or a letter. We suggest the letter “L” when you are in doubt as to what to use.

A left censored observation provides an upper bound for the actual failure time. All we know is that the failure time occurred at some point before the time value. Left censoring occurs when the items are not checked until some time after the study has begun. When a failed item is found, we do not know exactly when it failed, only that it was at some point before the left censor time.

Interval

When this value is entered in the Censor Variable, the corresponding time value is treated as an interval censored data value. The value may be a number or a letter. We suggest the letter “I” when you are in doubt as to what to use. When interval censoring is specified, the program uses both the Time Variable and the Start Time Variable.

An interval censored observation is one in which we know that the failure occurred between the two time values, but we do not know exactly when. This type of data is often called *readout* data. It occurs in situations where items are checked periodically for failures.

Note that when interval data are obtained, the first observation is usually left censored and the last observation is usually right censored.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Frequency Variable

Frequency Variable

This variable gives the count, or frequency, of the time displayed on that row. When omitted, each row receives a frequency of one. Frequency values should be positive integers. This is usually used to indicate the number of right censored values at the end of a study or the number of failures occurring within an interval. It may also be used to indicate ties for failure data.

Note that the probability plot estimation procedure requires that repeated time values be combined into a single observation which uses the value of this variable to represent the number of items.

Distribution to Fit

Distribution

This option specifies which probability distribution is fit. All results are for the specified probability distribution. If you select Find Best, the program displays reports and graphs that will help you select an appropriate probability distribution. When in the Find Best mode, the regular individual distribution reports and graphs are not displayed.

Distribution to Fit – Distributions Searched

Exponential - Weibull

These options are used by the distribution search procedure (Distribution = Find Best) to specify which of the available probability distributions should be included in the search.

Options

Threshold (Shift) Parameter

This option controls the setting and estimation of the threshold parameter. When this value is set to zero (which is the default) the threshold parameter is not fit. You can put in a fixed, nonzero value for the threshold here or you can specify ‘Search 0.’ Specifying a fixed value sets the threshold to that value. Specifying ‘Search 0’ causes a search for the threshold parameter to be conducted during the probability plot regression phase. The probability plot estimate is then used as if it were a fixed value in the maximum likelihood estimation phase.

In the case of the probability plot (least squares) estimates, a grid search is conducted for the value of the threshold that maximizes the correlation of the data on the probability plot.

Options Tab

The following options control the algorithms used during parameter estimation.

Estimation Options – Differential Evolution

Maximum Generations

Specify the maximum number of differential evolution iterations used to find a starting value before switching to Newton's method. A value between 100 and 200 is usually adequate. For large datasets (number of rows greater than 1000), you may want to reduce this number. Your strategy would be to let this algorithm provide reasonable starting values for Newton's algorithm which converges much more quickly when it has good starting values.

Individuals

This is the number of trial points that are used by the differential evolution algorithm at each iteration. In the terminology of differential evolution, this is the population size. A value between 15 and 25 is recommended. More points may dramatically increase the running time. Fewer points may not allow the algorithm to converge.

Inheritance

This value controls the amount of movement of the differential evolution algorithm toward the current best. Usually, a value between 0.5 and 1.0 is used. We suggest 0.85. A larger value accelerates movement toward the current best, but reduces the chance of locating the global maximum. A smaller value improves the chances of finding the global, rather than a local, solution, but increases the number of iterations until convergence.

Mutation Rate

This value controls the mutation rate of the differential evolution algorithm. This is the probability that the random adjustment of a parameter is set to zero—which is a *mutation* in the algorithm. Values between 0 and 1 are allowed. A value of 0.3 is recommended.

Grid Range

This is the initial range about each of the initial parameter values that is sampled during the differential evolution algorithm. The algorithm is not limited to this range, but specifying a value large enough to include the solution will increase the probability of convergence.

Estimation Options – Newton's Method

Maximum Iterations

This option assigns a maximum to the number of iterations used while performing Newton's method. We suggest a value of 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained.

Maximum Restarts

If Newton's method begins to diverge, it is restarted using slightly different starting values. This option specifies the number of times the algorithm is restarted.

Minimum Change

This value is used to terminate Newton's method. When the relative change in all of the parameters is less than this amount, Newton's method is terminated.

Step Adjustment

Newton's method calculates a change for each parameter value at each step. Instead of taking the whole parameter change, this option lets you take only a fraction of the indicated change. For datasets that diverge, taking only partial steps may allow the algorithm to converge. In essence, the algorithm tends to over correct the parameter values. This factor allows you to dampen this over correction. We suggest a value of about 0.2. This may increase the number of iterations (and you will have to increase the Max Iterations accordingly), but it provides a greater likelihood that the algorithm will converge.

Step Reduction

When Newton's method fails to converge, the Step Adjustment is reduced by multiplying by this amount. This forces Newton's method to take smaller steps which provides a better chance at convergence.

Estimation Options – Miscellaneous

Derivatives

This value specifies the machine precision value used in calculating numerical derivatives. Slight adjustments to this value can change the accuracy of the numerical derivatives (which impacts the variance/covariance matrix estimation).

Remember from calculus that the derivative is the slope calculated at a point along the function. It is the limit found by calculating the slope between two points on the function curve that are very close together. Numerical differentiation mimics this limit by calculating the slope between two function points that are very close together and then computing the slope. This value controls how close together these two function points are.

Numerical analysis suggests that this distance should be proportional to the machine precision of the computer. We have found that our algorithm achieves four-place accuracy in the variance-covariance matrix no matter what value is selected here (within reason). However, increasing or decreasing this value by two orders of magnitude may achieve six or seven place accuracy in the variance-covariance matrix. We have found no way to find the optimal value except trial and error.

Note that the parameter estimates do not seem to be influenced a great deal, only their standard errors.

Parameter 1 (Shape)

Specify a starting value for parameter one, the shape (location) parameter. If convergence fails, try a different value here or try increasing the grid range. Select 'Data' to calculate an appropriate value from the data.

Parameter 2 (Scale)

Specify a starting value for parameter two, the scale parameter. If convergence fails, try a different value here or try increasing the grid range. Select 'Data' to calculate an appropriate value from the data.

Prob Plot Model

When a probability plot method is used to estimate the parameters of the probability distribution, this option designates which variable (time or frequency) is used as the dependent variable.

- **F=A+B(Time)**

  On the probability plot, F is regressed on Time and the resulting intercept and slope are used to estimate the parameters. See the discussion of probability plots below for more information.

- **Time=A+B(F)**

  On the probability plot, Time is regressed on F and the resulting intercept and slope are used to estimate the parameters.

Hazard Options

The following options control the calculation of the hazard rate and cumulative hazard function.

Bandwidth Method

This option designates the method used to specify the smoothing bandwidth used to calculate the hazard rate. Specify an amount or a percentage of the time range. The default is to specify a percent of the time range.

Bandwidth Amount

This option specifies the bandwidth size used to calculate the hazard rate. If the Bandwidth Method was set to Amount, this is a value in time units (such as 10 hours). If Percentage of Time Range was selected, this is a percentage of the overall range of the data.

Smoothing Kernel

This option specifies the kernel function used in the smoothing to estimate the hazard rate. You can select *uniform*, *Epanechnikov*, or *biweight* smoothing. The actual formulas for these functions are provided later in the Hazard Rate output section.

Product Limit and Hazard Confidence Limits Method

The standard nonparametric estimator of the reliability function is the Product Limit estimator. This option controls the method used to estimate the confidence limits of the estimated reliability. The options are Linear, Log Hazard, Arcsine Square Root, and Nelson-Aalen. The formulas used by these options were presented earlier. Although the Linear (Greenwood) is the most commonly used, recent studies have shown that either the Log Hazard or the Arsine Square Root Hazard are better in the sense that they require a smaller sample size to be accurate. The Nelson-Aalen has also become a more popular choice.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report - Parametric Hazard Rate Report

These options indicate whether to display the corresponding report.

Report Options

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Survival and Haz Rt Calculation Values

Percentiles

This option specifies a list of percentiles (range 1 to 99) at which the reliability (survivorship) is reported. The values should be separated by commas.

Specify sequences with a colon, putting the increment inside parentheses after the maximum in the sequence. For example: 5:25(5) means 5,10,15,20,25 and 1:5(2),10:20(2) means 1,3,5,10,12,14,16,18,20.

Times

This option specifies a list of times at which the percent surviving is reported. Individual values are separated by commas. You can specify a sequence by specifying the minimum and maximum separate by a colon and putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers. Use ‘(10)’ alone to specify ten values between zero and the maximum value found in the data.

Report Options – Residual Life Calculation Values

Residual Life Percentiles

This option specifies a list of up to four percentiles (range 1 to 99) at which the residual life is reported. The values should be separated by commas. Only the first four values are used on the report.

Report Options – Decimal Places

Time

This option specifies the number of decimal places shown on reported time values.

Probability

This option specifies the number of decimal places shown on reported probability and hazard values.

Plots Tab

The following options control which plots are displayed.

Select Plots

Survival/Reliability Plot - Probability Plot

These options specify which plots are displayed.

Select Plots – Plots Displayed

Show Individual Plots

When checked, this option specifies that a separate chart is to be displayed for each group (as specified by the Group Variable).

Show Combined Plots

When checked, this option specifies that a chart combining all groups is to be displayed.

Plot Options

Number of Intervals

This option specifies the number of points along the time axis at which calculations are made. This controls the resolution of the plots. Usually, a value between 50 and 100 is sufficient.

Plot Options – Plot Arrangement

Two Plots Per Line

When a lot of charts are specified, checking this option will cause the size of the charts to be reduced so that they can be displayed two per line. This will reduce the overall size of the output.

Plot Options – Plot Contents

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the name of the group variable.

Survival Plot Tab

These options control the attributes of the survival (reliability) curves.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Survival Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Censor Tickmarks

This option indicates the size of the tickmarks (if any) showing where the censored points fall on the Kaplan-Meier survival curve. The values are at a scale of 1000 = one inch.

RECOMMENDED: Enter 0 for no censor tickmarks. Enter 100 when you want to display the tickmarks.

Survival Plot Settings – Plot Contents

Nonparametric Survival

This box indicates whether to display the Kaplan-Meier survival curve.

Nonparametric Confidence Limits

This box indicates whether to display confidence limits about the nonparametric survival curve.

Parametric Survival

This box indicates whether to display the parametric survival curve, based on the selected probability distribution.

Parametric Confidence Limits

This box indicates whether to display confidence limits about the parametric survival curve.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, *{Z}*, and *{M}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Cum Haz Plot Tab

These options control the attributes of the cumulative hazard function plot.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Cum Haz Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Cum Haz Plot Settings – Plot Contents

Nonparametric Hazard Function

This box indicates whether to display the Kaplan-Meier hazard curve.

Parametric Hazard Function

This box indicates whether to display the parametric cumulative hazard curve, based on the probability distribution that was selected.

Parametric Hazard Confidence Limits

This box indicates whether to display confidence limits about the parametric hazard curve.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}, {X}, {Z},* and *{M}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Haz Rt Plot Tab

These options control the attributes of the hazard rate plot.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Haz Rt Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Haz Rt Plot Settings – Plot Contents

Nonparametric Hazard Rate

This box indicates whether to display the smoothed, nonparametric hazard rate curve.

Parametric Hazard Rate

This box indicates whether to display the parametric hazard rate curve, based on the probability distribution that was selected.

Parametric Hazard Rate Confidence Limits

This box indicates whether to display confidence limits about the parametric hazard rate. Note that these are asymptotic confidence limits based on large sample results. We have found that these limits do not work well for some distributions, especially the lognormal and the log-logistic.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, *{Z}*, and *{M}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

These options control the attributes of the probability plot(s).

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Prob Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Let us emphasize that this probability plot uses the Scatter Plot style files, not the Probability Plot style files.

Plotting Position - F(T)

The probability plot shows either time or the log of time on the vertical axis (depending on the distribution) and the appropriate transformation of F(t) on the horizontal axis. Note that F(t) is the cumulative distribution function. This option specifies the method used to determine F(t)—used to calculate the vertical plotting positions on the probability plot. Note that method selected here also influences the probability plot estimates of the parameters.

The five alternatives available are

- **Median (j-0.3)/(n+0.4)**

  The most popular method is to calculate the median rank for each sorted data value. That is, this is the value for the j^{th} sorted time value. Since the median rank requires extensive calculations, this approximation to the median rank is often used.

$$F(t_j) = \frac{j - 0.3}{n + 0.4}$$

- **Median (Exact)**

  The most popular method is to calculate the median rank for each data value. This is the median rank of the j^{th} sorted time value out of n values. This method will be a little more time consuming to calculate. The exact value of the median rank is calculated using the formula

$$F(t_j) = \frac{1}{1 + \frac{n - j + 1}{j} F_{0.50;\, 2(n-j+1);\, 2j}}$$

- **Mean j/(n+1)**
  The mean rank is sometimes recommended. In this case, the formula is

$$F(t_j) = \frac{j}{n+1}$$

- **White (j-3/8)/(n+1/4)**
  A formula proposed by White is sometimes recommended. The formula is

$$F(t_j) = \frac{j+3/8}{n+1/4}$$

- **(j-0.5)/n**
  The following formula is sometimes used

$$F(t_j) = \frac{j-0.5}{n}$$

Prob Plot Settings – Plot Contents

Trend Line

This option controls whether the trend (least squares) line is calculated and displayed.

Residuals from Trend Line

This option controls whether the vertical deviations from the trend line are displayed. Displaying these residuals may let you see departures from linearity more easily.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, *{Z}*, and *{M}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines & Symbols Tab

These options specify the attributes of the lines used for each group in the hazard curves and survival curves and the symbols used for each group in the probability plots.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Plotting Symbols

Symbol 1 - 15

These options specify the symbols used in the plot of each group. The first symbol is used by the first group, the second symbol by the second group, and so on. These symbols are provided to allow the various groups to be easily identified, even on black and white printers.

Clicking on a symbol box (or the small button to the right of the symbol box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Labels Tab

The options on this tab specify the labels that a printed on the reports and plots.

Report and Plot Labels

Failure Time Label - Residual Life Label

These options specify the term(s) used as headings for these items in the reports and labels on the plots. Since these reports are used for performing survival analysis in medical research and reliability analysis in industry, and since these fields often use different terminology, these options are needed to provide appropriate headings for the reports.

Storage Tab

These options control the storage of information back to the database for further use.

Data Storage Variables

Failure Time - Parametric UCL

Each of the fields on these two options let you specify columns (variables) on the database to which the corresponding data are automatically stored.

Warning: existing data are replaced, so make sure that the columns you select are empty.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Weibull Distribution

This section presents an example of how to fit the Weibull distribution. The data used were shown above and are found in the WEIBULL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Distribution (Weibull) Fitting window.

1 Open the WEIBULL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **WEIBULL.S0**.
- Click **Open**.

2 Open the Distribution (Weibull) Fitting window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Distribution (Weibull) Fitting**. The Distribution (Weibull) Fitting procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Distribution (Weibull) Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Time** from the list of variables and then click **Ok**.
- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **Censor** from the list of variables and then click **Ok**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**.

4 Set the Derivative constant.

- On the Distribution (Weibull) Fitting window, select the **Options tab**.
- Set the **Derivatives** value to **0.00006**.

5 Specify the plots.

- On the Distribution (Weibull) Fitting window, select the **Plots tab**.
- In addition to the items that are already checked, check **Hazard Function Plot** and **Hazard Rate Plot**.
- On the Distribution (Weibull) Fitting window, select the **Survival Plot tab**.
- In addition to the items that are already checked, check **Parametric C.L.**

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum
Failed	12	12	12.5	152.7
Right Censored	1	18	152.7	152.7
Left Censored	0	0		
Interval Censored	0	0		
Total	13	30	12.5	152.7

This report displays a summary of the amount of data that were analyzed. Scan this report to determine if there were any obvious data errors by double checking the counts and the minimum and maximum.

Parameter Estimation Section

Weibull Parameter Estimation Section

Parameter	Probability Plot Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Shape	1.26829	1.511543	0.4128574	0.8849655	2.581753
Scale	279.7478	238.3481	57.21616	148.8944	381.5444
Threshold	0	0			
Log Likelihood		-80.05649			
Mean	259.7101	214.9709			
Median	209.5383	187.0276			
Mode	82.19741	116.3898			
Sigma	206.2112	144.9314			
Differential Evolution Iterations		31			
Newton Raphson Restart		1			
Newton Raphson Iterations		8			

This report displays parameter estimates along with standard errors and confidence limits for the maximum likelihood estimates. In this example, we have set the threshold parameter to zero so we are fitting the two-parameter Weibull distribution.

Probability Plot Estimate

This estimation procedure uses the data from the probability plot to estimate the parameters. The estimation formula depends on which option was selected for the Prob Plot Model (in the Estimation tab window).

Prob Plot Model: F=A+B(Time)

The cumulative distribution function *F(t)*

$$F(t) = 1 - e^{-\left(\frac{t-D}{C}\right)^B}$$

may be rearranged as (assuming *D* is zero)

$$\ln(-\ln(1-F(t))) = -B\left[\ln(C)\right] + B\left[\ln(t)\right]$$

This is now in a linear form. If we let *y = ln(-ln(1-F(t)))* and *x = ln(t)*, the above equation becomes

$$y = -B\left[\ln(C)\right] + Bx$$

Using simple linear regression, we can estimate the intercept and slope. Using these estimates, we obtain estimates of the Weibull parameters *B* and *C* as

$$B = \text{slope}$$

and

$$C = \exp\left(\frac{-\text{intercept}}{B}\right)$$

We assumed that *D* was zero. If *D* is not zero, it is treated as a known value and subtracted from the time values before the above regression is calculated.

Prob Plot Model: Time=A+B(F)

The cumulative distribution function *F(t)*

$$F(t) = 1 - e^{-\left(\frac{t-D}{C}\right)^B}$$

may be rearranged as (assuming *D* is zero)

$$\left(\frac{1}{B}\right)\ln(-\ln(1-F(t))) + \ln(C) = \ln(t)$$

This is now in a linear form. If we let *x = ln(-ln(1-F(t)))* and *y = ln(t)*, the above equation becomes

$$y = \frac{1}{B}x + \ln(C)$$

Using simple linear regression, we can estimate the intercept and slope. Using these estimates, we obtain estimates of the Weibull parameters *B* and *C* as

$$B = \frac{1}{\text{slope}}$$

and

$$C = \exp\left(\text{intercept}\right)$$

Parameter estimates for the other distributions are found in a similar manner.

Maximum Likelihood Estimates of B, C, M, and S

These are the usual maximum likelihood estimates (MLE) of the parameters. The formulas for the standard errors and confidence limits use the estimated variance covariance matrix which is the inverse of the Fisher information matrix, $\left\{vc_{i,j}\right\}$. The standard errors are given as the square roots of the diagonal elements $vc_{1,1}$ and $vc_{2,2}$.

In the case of the Weibull distribution, the confidence limits for *B* are

$$\hat{B}_{lower,1-\alpha/2} = \frac{\hat{B}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{1,1}}}{\hat{B}}\right\}}$$

$$\hat{B}_{upper,1-\alpha/2} = \hat{B}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{1,1}}}{\hat{B}}\right\}$$

In the case of the Weibull distribution, the confidence limits for *C* are

$$\hat{C}_{lower,1-\alpha/2} = \frac{\hat{C}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{C}}\right\}}$$

$$\hat{C}_{upper,1-\alpha/2} = \hat{C}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{C}}\right\}$$

In the case of all other distributions, the confidence limits for *M* are

$$\hat{M}_{lower,1-\alpha/2} = \hat{M} - z_{1-\alpha/2}\sqrt{vc_{1,1}}$$

$$\hat{M}_{upper,1-\alpha/2} = \hat{M} + z_{1-\alpha/2}\sqrt{vc_{1,1}}$$

In the case of all other distributions, the confidence limits for *S* are

$$\hat{S}_{lower,1-\alpha/2} = \frac{\hat{S}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{S}}\right\}}$$

$$\hat{S}_{upper,1-\alpha/2} = \hat{S}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{S}}\right\}$$

$$\hat{S}_{upper,1-\alpha/2} = \hat{S}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{S}}\right\}$$

Log Likelihood

This is the value of the log likelihood function calculated using the maximum likelihood parameter estimates. This is the value of the function being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood values achieved by each distribution and select as the best fitting the one with the maximum value.

Note that we have found that several popular statistical programs calculate this value without including all of the terms. Hence, they present erroneous values. The error occurs because they omit the 1/*t* term in the denominator of the lognormal and the Weibull log-likelihood functions.

Also, they may fail to include a correction for using the logarithm to the base 10 in the lognormal10. Hopefully, future editions of these programs will calculate the likelihood correctly.

Mean

This is the mean time to failure (MTTF). It is the mean of the random variable (failure time) being studied given that the fitted distribution provides a reasonable approximation to your data's actual distribution. In the case of the Weibull distribution, the formula for the mean is

$$Mean = D + C\,\Gamma\left(1 + \frac{1}{B}\right)$$

where $\Gamma(x)$ is the gamma function.

Median

The median is the value of *t* where *F(t)=0.5*. In the case of the Weibull distribution, the formula for the median is

$$Median = D + C(\log 2)^{1/B}$$

Mode

The mode of the Weibull distribution is given by

$$Mode = D + C\left(1 - \frac{1}{B}\right)^{1/B}$$

Sigma

This is the standard deviation of the failure time. The formula for the standard deviation (sigma) of a Weibull random variable is

$$\sigma = C\sqrt{\Gamma\left(1 + \frac{2}{B}\right) - \Gamma^2\left(1 + \frac{1}{B}\right)}$$

where $\Gamma(x)$ is the gamma function.

Differential Evolution Iterations

This is the number of iterations used in the differential evolution phase of the maximum likelihood algorithm. If this value is equal to the maximum number of generations allowed, the algorithm did not converge, so you should increase the maximum number of generations and re-run the procedure.

Newton Raphson Restarts

This is the number of times the Newton Raphson phase of the maximum likelihood algorithm was restarted because the algorithm diverged. Make sure that the maximum number of restarts was not reached.

Newton Raphson Iterations

This is the number of iterations used in the Newton Raphson phase of the maximum likelihood algorithm. If this value is equal to the maximum number of iterations allowed, the algorithm did not converge, so you should increase the maximum number of iterations and re-run the procedure.

Inverse of Fisher Information Matrix

Inverse of Fisher Information Matrix

Parameter	Shape	Scale
Shape	0.1706034	-14.33367
Scale	-14. 33367	3273.362

This table gives the inverse of the Fisher information matrix evaluated at the maximum likelihood estimates which is an asymptotic estimate of the variance-covariance matrix of the two parameters. These values are calculated using numerical second-order derivatives.

Note that because these are numerical derivatives based on a random start provided by differential evolution, the values of the last two decimal places may vary from run to run. You can stabilize the values by changing the value of Derivatives constant, but this will have little effect on the overall accuracy of your results.

Kaplan-Meier Product-Limit Survival Distribution

Kaplan-Meier Product-Limit Survival Distribution

Failure Time	Survival	Lower 95% C.L.	Upper 95% C.L.	Hazard Fn	Lower 95% C.L.	Upper 95% C.L.	Sample Size
12.5	0.9667	0.9024	1.0000	0.0339	0.0000	0.1027	30
24.4	0.9333	0.8441	1.0000	0.0690	0.0000	0.1695	29
58.2	0.9000	0.7926	1.0000	0.1054	0.0000	0.2324	28
68.0	0.8667	0.7450	0.9883	0.1431	0.0118	0.2943	27
69.1	0.8333	0.7000	0.9667	0.1823	0.0339	0.3567	26
95.5	0.8000	0.6569	0.9431	0.2231	0.0585	0.4203	25
96.6	0.7667	0.6153	0.9180	0.2657	0.0855	0.4856	24
97.0	0.7333	0.5751	0.8916	0.3102	0.1148	0.5532	23
114.2	0.7000	0.5360	0.8640	0.3567	0.1462	0.6236	22
123.2	0.6667	0.4980	0.8354	0.4055	0.1799	0.6972	21
125.6	0.6333	0.4609	0.8058	0.4568	0.2160	0.7746	20
152.7	0.6000	0.4247	0.7753	0.5108	0.2545	0.8564	19
152.7+							18

Confidence Limits Method: Linear (Greenwood)

This report displays the Kaplan-Meier product-limit survival distribution and hazard function along with confidence limits. The formulas used were presented earlier. Note that these estimates do not use the selected parametric distribution in any way. They are the nonparametric estimates and are completely independent of the distribution that is being fit.

Note that censored observations are marked with a plus sign on their time value. The survival and hazard functions are not calculated for censored observations. Also note that left censored and interval censored observations are treated as failed observations for the calculations on this report.

Also note that the Sample Size is given for each time period. As time progresses, participants are removed from the study, reducing the sample size. Hence, the survival results near the end of the study are based on only a few participants and are therefore less precise. This shows up as a widening of the confidence limits.

Nonparametric Hazard Rate Section

Nonparametric Hazard Rate Section

Failure Time	Nonparametric Hazard Rate	Std Error of Hazard Rate	95% Lower Conf. Limit of Hazard Rate	95% Upper Conf. Limit of Hazard Rate
8.0	0.0016	0.0014	0.0003	0.0090
16.0	0.0019	0.0014	0.0005	0.0078
24.0	0.0016	0.0012	0.0004	0.0066
32.0	0.0015	0.0010	0.0004	0.0052
40.0	0.0016	0.0009	0.0005	0.0047
48.0	0.0021	0.0011	0.0008	0.0059
56.0	0.0024	0.0014	0.0008	0.0075
64.0	0.0027	0.0015	0.0009	0.0082
72.0	0.0036	0.0016	0.0015	0.0086
80.0	0.0042	0.0017	0.0019	0.0095
88.0	0.0043	0.0018	0.0018	0.0099
96.0	0.0045	0.0019	0.0019	0.0105
104.0	0.0052	0.0022	0.0023	0.0117
112.0	0.0054	0.0022	0.0024	0.0121
120.0	0.0047	0.0021	0.0019	0.0113
128.0	0.0038	0.0020	0.0014	0.0105
136.0	0.0038	0.0021	0.0013	0.0111
144.0	0.0038	0.0036	0.0006	0.0246
152.0	0.0081	0.0081	0.0011	0.0575

This report displays the nonparametric estimate of the hazard rate, *h(t)*. Note that this is not the cumulative hazard function *H(t)* shown in the last report. It is the derivative of *H(t)*. Since it is *h(t)* that needs to be studied in order to determine the characteristics of the failure process, this report and its associated plot (which is shown below) become very import.

The formula for the Nelson-Aalen estimator of the cumulative hazard is

$$\tilde{H}(t) = \begin{cases} 0 & if\ \ t < t_1 \\ \sum_{t_1 \le t} \frac{d_i}{Y_i} & if\ \ t_1 \le t \end{cases}$$

The variance of this estimate is

$$\sigma_H^2(t) = \sum_{t_i \le t} \frac{(Y_i - d_i)d_i}{(Y_i - 1)Y_i^2}$$

In the above equation, d_i represents the number of deaths at time t_i and Y_i represents the number of individuals who are at risk at time t_i.

The hazard rate is estimated using kernel smoothing of the Nelson-Aalen estimator as given in Klein and Moeschberger (1997). The formulas for the estimated hazard rate and its variance are given by

$$\hat{h}(t) = \frac{1}{b}\sum_D K\left(\frac{t - t_k}{b}\right)\Delta\tilde{H}(t_k)$$

$$\sigma^2\left[\hat{h}(t)\right] = \frac{1}{b^2}\sum_D K\left(\frac{t - t_k}{b}\right)^2 \Delta\hat{V}\left[\tilde{H}(t_k)\right]$$

where *b* is the bandwidth about *t* and

$$\Delta\tilde{H}(t_k) = \tilde{H}(t_k) - \tilde{H}(t_{k-1})$$

$$\Delta\hat{V}\left[\tilde{H}(t_k)\right] = \hat{V}\left[\tilde{H}(t_k)\right] - \hat{V}\left[\tilde{H}(t_{k-1})\right]$$

Three choices are available for the kernel function *K(x)* in the above formulation. These are defined differently for various values of *t*. Note that the t_k's are for failed items only and that t_D is the maximum failure time. For the *uniform kernel* the formulas for the various values of *t* are

$$K(x) = \frac{1}{2} \quad for \quad t - b \le t \le t + b$$

$$K_L(x) = \frac{4\left(1+q^3\right)}{(1+q)^4} + \frac{6(1-q)}{(1+q)^3}x \quad for \quad t < b$$

$$K_R(x) = \frac{4\left(1+r^3\right)}{(1+r)^4} - \frac{6(1-r)}{(1+r)^3}x \quad for \quad t_D - b < t < t_D$$

where

$$q = \frac{t}{b}$$

and

$$r = \frac{t_D - t}{b}$$

For the *Epanechnikov kernel* the formulas for the various values of *t* are

$$K(x) = \frac{3}{4}\left(1-x^2\right) \quad for \quad t - b \le t \le t + b$$

$$K_L(x) = K(x)(A + Bx) \quad for \quad t < b$$

$$K_R(x) = \frac{4\left(1+r^3\right)}{(1+r)^4} - \frac{6(1-r)}{(1+r)^3}x \quad for \quad t_D - b < t < t_D$$

where

$$A = \frac{64\left(2 - 4q + 6q^2 - 3q^3\right)}{(1+q)^4\left(19 - 18q + 3q^2\right)}$$

$$B = \frac{240(1-q)^2}{(1+q)^4\left(19 - 18q + 3q^2\right)}$$

$$q = \frac{t}{b}$$

$$r = \frac{t_D - t}{b}$$

For the *biweight kernel* the formulas for the various values of *t* are

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \quad for \quad t - b \le t \le t + b$$

$$K_L(x) = K(x)(A + Bx) \quad for \quad t < b$$

$$K_R(x) = K(-x)(A - Bx) \quad for \quad t_D - b < t < t_D$$

where

$$A = \frac{64\left(8 - 24q + 48q^2 - 45q^3 + 15q^4\right)}{(1+q)^5\left(81 - 168q + 126q^2 - 40q^3 + 5q^4\right)}$$

$$B = \frac{1120(1-q)^3}{(1+q)^5\left(81 - 168q + 126q^2 - 40q^3 + 5q^4\right)}$$

$$q = \frac{t}{b}$$

$$r = \frac{t_D - t}{b}$$

Confidence intervals for *h(t)* are given by

$$\hat{h}(t)\exp\left[\pm \frac{z_{1-\alpha/2}\sigma\left[\hat{h}(t)\right]}{\hat{h}(t)}\right]$$

Care must be taken when using these kernel-smoothed estimators since they are actually estimating a smoothed version of the hazard rate, not the hazard rate itself. Thus, they may be biased and are greatly influenced by the choice of the bandwidth *b*. We have found that you must experiment with *b* to find an appropriate value for each dataset.

Parametric Hazard Rate Section

Weibull Hazard Rate Section

Failure Time	Weibull Hazard Rate	95% Lower Conf. Limit of Hazard Rate	95% Upper Conf. Limit of Hazard Rate
8.0	0.0011	0.0005	0.0023
16.0	0.0016	0.0008	0.0032
24.0	0.0020	0.0010	0.0040
32.0	0.0023	0.0011	0.0046
40.0	0.0025	0.0012	0.0052
48.0	0.0028	0.0014	0.0057
56.0	0.0030	0.0015	0.0062
64.0	0.0032	0.0016	0.0066
72.0	0.0034	0.0017	0.0070
80.0	0.0036	0.0018	0.0074
88.0	0.0038	0.0019	0.0078
96.0	0.0040	0.0020	0.0081
104.0	0.0041	0.0020	0.0085
112.0	0.0043	0.0021	0.0088
120.0	0.0045	0.0022	0.0091
128.0	0.0046	0.0023	0.0094
136.0	0.0048	0.0023	0.0097
144.0	0.0049	0.0024	0.0100
152.0	0.0050	0.0025	0.0103
160.0	0.0052	0.0025	0.0105

This report displays the maximum likelihood estimates of the hazard rate, *h(t)*, based on the selected probability distribution and the definition of the hazard rate

$$h(t) = \frac{f(t)}{R(t)}$$

Asymptotic confidence limits are computed using the formula from Nelson (1991) page 294.

$$\hat{h}(t)\exp\left[\pm\frac{z_{1-\alpha/2}s\left[\hat{h}(t)\right]}{\hat{h}(t)}\right]$$

where

$$s^2\left[\hat{h}(t)\right] = \left(\frac{\partial\hat{h}}{\partial P_1}\right)^2 vc_{1,1} + \left(\frac{\partial\hat{h}}{\partial P_2}\right)^2 vc_{2,2} + 2\left(\frac{\partial\hat{h}}{\partial P_1}\right)\left(\frac{\partial\hat{h}}{\partial P_2}\right) vc_{1,2}$$

The partial derivatives are evaluated using numerical differentiation.

Note that we have found that the above approximation behaves poorly for some distributions. However, this is the only formula that we have been able to find, so this is what we provide. If you find that the confidence limits have a strange appearance (especially, in that the width goes to zero), please ignore them. They should appear as nice expanding lines about the estimated hazard rate.

Parametric Failure Distribution Section

Weibull Failure Distribution Section

Failure Time	Prob Plot Estimate of Failure	Max Like Estimate of Failure	95% Lower Conf. Limit of Failure	95% Upper Conf. Limit of Failure
8.0	0.0110	0.0059	0.0005	0.0622
16.0	0.0262	0.0167	0.0027	0.1011
24.0	0.0434	0.0306	0.0067	0.1345
32.0	0.0619	0.0469	0.0127	0.1649
40.0	0.0814	0.0651	0.0209	0.1935
48.0	0.1014	0.0849	0.0310	0.2209
56.0	0.1219	0.1060	0.0431	0.2477
64.0	0.1427	0.1281	0.0569	0.2741
72.0	0.1637	0.1510	0.0723	0.3005
80.0	0.1849	0.1747	0.0888	0.3272
88.0	0.2060	0.1989	0.1064	0.3543
96.0	0.2271	0.2235	0.1245	0.3819
104.0	0.2480	0.2483	0.1431	0.4102
112.0	0.2689	0.2734	0.1617	0.4391
120.0	0.2895	0.2984	0.1801	0.4688
128.0	0.3099	0.3234	0.1982	0.4991
136.0	0.3301	0.3483	0.2158	0.5298
144.0	0.3500	0.3730	0.2328	0.5607
152.0	0.3695	0.3975	0.2491	0.5917
160.0	0.3888	0.4216	0.2648	0.6225

This report displays the estimated values of the cumulative failure distribution, *F(t)*, at the time values that were specified in the Times option of the Reports Tab. These failure values are the estimated probability that failure occurs by the given time point. For example, the maximum likelihood estimate that a unit will fail within 88 hours is 0.1989. The 95% confidence estimate of this probability is 0.1064 to 0.3543.

The asymptotic confidence limits are computed using the following formula:

$$\hat{F}_L(t) = \hat{F}\left(\hat{u} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

$$\hat{F}_U(t) = \hat{F}\left(\hat{u} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

where

$$\hat{u} = \frac{t - \hat{\mu}}{\hat{\sigma}}$$

$$\hat{V}(\hat{u}) = \frac{V(\hat{\mu}) + \hat{u}^2 V(\hat{\sigma}) + 2\hat{u}Cov(\hat{\mu},\hat{\sigma})}{\hat{\sigma}^2}$$

Note that limits for the Weibull, lognormal, and log-logistic are found using the corresponding extreme value, normal, and logistic probability functions using the substitution *y=ln(t)*.

Parametric Reliability Section

Reliability Section

Failure Time	Prob Plot Estimate of Survival	Max Like Estimate of Survival	95% Lower Conf. Limit of Survival	95% Upper Conf. Limit of Survival
8.0	0.9890	0.9941	0.9378	0.9995
16.0	0.9738	0.9833	0.8989	0.9973
24.0	0.9566	0.9694	0.8655	0.9933
32.0	0.9381	0.9531	0.8351	0.9873
40.0	0.9186	0.9349	0.8065	0.9791
48.0	0.8986	0.9151	0.7791	0.9690
56.0	0.8781	0.8940	0.7523	0.9569
64.0	0.8573	0.8719	0.7259	0.9431
72.0	0.8363	0.8490	0.6995	0.9277
80.0	0.8151	0.8253	0.6728	0.9112
88.0	0.7940	0.8011	0.6457	0.8936
96.0	0.7729	0.7765	0.6181	0.8755
104.0	0.7520	0.7517	0.5898	0.8569
112.0	0.7311	0.7266	0.5609	0.8383
120.0	0.7105	0.7016	0.5312	0.8199
128.0	0.6901	0.6766	0.5009	0.8018
136.0	0.6699	0.6517	0.4702	0.7842
144.0	0.6500	0.6270	0.4393	0.7672
152.0	0.6305	0.6025	0.4083	0.7509
160.0	0.6112	0.5784	0.3775	0.7352

This report displays the estimated reliability (survival) at the time values that were specified in the Times option of the Reports Tab. Reliability may be thought of as the probability that failure occurs after the given failure time. Thus, (using the ML estimates) the probability is 0.9531 that failure will not occur until after 32 hours. The 95% confidence for this estimated probability is 0.8351 to 0.9873.

Two reliability estimates are provided. The first uses the parameters estimated from the probability plot and the second uses the maximum likelihood estimates. Confidence limits are calculated for the maximum likelihood estimates. (They have not been derived for the probability plot estimates for all data situations). The formulas used are as follows.

$$\hat{R}_L(t) = \hat{R}\left(\hat{u} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

$$\hat{R}_U(t) = \hat{R}\left(\hat{u} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

where

$$\hat{u} = \frac{t - \hat{M}}{\hat{S}}$$

$$\hat{V}(\hat{u}) = \frac{V\left(\hat{M}\right) + \hat{u}^2 V\left(\hat{S}\right) + 2\hat{u}Cov\left(\hat{M}, \hat{S}\right)}{\hat{S}^2}$$

Note that limits for the Weibull, lognormal, and log-logistic are found using the corresponding extreme value, normal, and logistic probability functions using the substitution *y=ln(t)*.

Parametric Percentile Section

Percentile Section

Failure Time Percentage	Prob Plot Estimate of Failure Time	Max Like Estimate of Failure Time	95% Lower Conf. Limit of Failure Time	95% Upper Conf. Limit of Failure Time
5.0000	26.9	33.4	14.2	78.4
10.0000	47.4	53.8	28.5	101.4
15.0000	66.8	71.6	42.7	120.3
20.0000	85.7	88.4	56.5	138.3
25.0000	104.7	104.5	69.7	156.8
30.0000	124.1	120.5	82.2	176.7
35.0000	144.0	136.5	93.9	198.6
40.0000	164.7	152.8	104.8	222.9
45.0000	186.5	169.6	115.0	250.1
50.0000	209.5	187.0	124.7	280.5
55.0000	234.3	205.4	134.0	314.8
60.0000	261.1	225.0	143.1	353.7
65.0000	290.7	246.1	152.1	398.3
70.0000	323.8	269.5	161.3	450.3
75.0000	361.9	295.8	170.9	512.1
80.0000	407.1	326.5	181.3	588.2
85.0000	463.5	364.1	193.0	686.7
90.0000	540.0	413.9	207.4	826.0
95.0000	664.5	492.6	227.8	1065.0

This report displays failure time percentiles and, for the maximum likelihood estimates, confidence intervals for those percentiles. For example, the estimated median failure time is 187 hours. The 95% confidence limits for the median time are 124.7 to 280.5. Note that these limits are very wide for two reasons. First, the sample size is small. Second, the shape parameter is less than 2.0.

The estimated 100pth percentile and associated confidence interval is computed using the following steps.

1. Compute $w_p = F^{-1}(p)$
2. Compute $y_p = \hat{M} + w_p \hat{S}$. Note that in the case of the Weibull and exponential distributions, we let $\hat{M} = \ln(\hat{C})$ and $\hat{S} = 1/\hat{B}$.
3. Compute $V(y_p) = VC_{1,1} + y_p^2 VC_{2,2} + 2 y_p VC_{1,2}$.
4. For the normal, extreme value, and logistic distributions, the confidence interval for the percentile is given by

$$T_{Lower,p} = y_p - z_{1-\alpha/2}\sqrt{V(y_p)} + D$$

$$T_{Upper,p} = y_p + z_{1-\alpha/2}\sqrt{V(y_p)} + D$$

For the lognormal, exponential, Weibull, and log-logistic distributions, the confidence interval for the percentile is given by

$$T_{Lower,p} = \exp\left(y_p - z_{1-\alpha/2}\sqrt{V\left(y_p\right)}\right) + D$$

$$T_{Upper,p} = \exp\left(y_p + z_{1-\alpha/2}\sqrt{V\left(y_p\right)}\right) + D$$

For the lognormal base 10 distribution, the confidence interval for the percentile is given by

$$T_{Lower,p} = 10^{y_p - z_{1-\alpha/2}\sqrt{V\left(y_p\right)}} + D$$

$$T_{Upper,p} = 10^{y_p + z_{1-\alpha/2}\sqrt{V\left(y_p\right)}} + D$$

Parametric Residual Life Section

Weibull Residual Life Section

Failure Time	25.0th %tile Proportion Failing	50.0th %tile Residual Life	75.0th %tile Residual Life	95.0th %tile Residual Life	Residual Life
8.0	0.0059	97.9	180.1	288.7	406.6
16.0	0.0167	92.5	174.0	282.2	399.9
24.0	0.0306	87.9	168.5	276.2	393.5
32.0	0.0469	83.8	163.5	270.6	387.5
40.0	0.0651	80.1	158.9	265.3	381.8
48.0	0.0849	76.9	154.5	260.2	376.3
56.0	0.1060	73.9	150.5	255.4	371.1
64.0	0.1281	71.3	146.7	250.9	366.0
72.0	0.1510	68.8	143.2	246.5	361.1
80.0	0.1747	66.6	139.9	242.4	356.4
88.0	0.1989	64.6	136.7	238.4	351.8
96.0	0.2235	62.7	133.8	234.5	347.4
104.0	0.2483	60.9	131.0	230.9	343.1
112.0	0.2734	59.3	128.3	227.3	339.0
120.0	0.2984	57.8	125.8	223.9	335.0
128.0	0.3234	56.4	123.4	220.7	331.1
136.0	0.3483	55.1	121.1	217.5	327.3
144.0	0.3730	53.8	118.9	214.5	323.6
152.0	0.3975	52.7	116.9	211.5	320.0
160.0	0.4216	51.6	114.9	208.7	316.6

This report gives percentiles of the estimating life remaining after a certain time period. For example, the estimated median remaining life of items reaching 80.0 hours is 139.9 hours.

The percentile and associated confidence interval of residual (remaining) life is computed using the following steps.

1. Compute $z_p = \dfrac{y_p - \hat{M}}{\hat{S}}$. Note that in the case of the Weibull and exponential distributions, we let $\hat{M} = \ln(\hat{C})$ and $\hat{S} = 1/\hat{B}$. Also note that for the normal, extreme value, and logistic distributions, $y_p = t$. For the lognormal, Weibull, and log-logistic distributions, $y_p = e^t$. And for the lognormal base 10 distribution, $y_p = 10^t$.
2. Compute $p_0 = F\left(z_p\right)$
3. Compute $p_1 = p_0\left(1 + P\right)$, where P is the percentile of residual life to be estimated.

4. Compute $w_p = \hat{M} + F^{-1}(p_1)\hat{S}$. Note that in the case of the Weibull and exponential distributions, we let $\hat{M} = \ln(\hat{C})$ and $\hat{S} = 1/\hat{B}$.

   For the normal, extreme value, and logistic distributions, the estimate is given by

   $$T_p = w_p$$

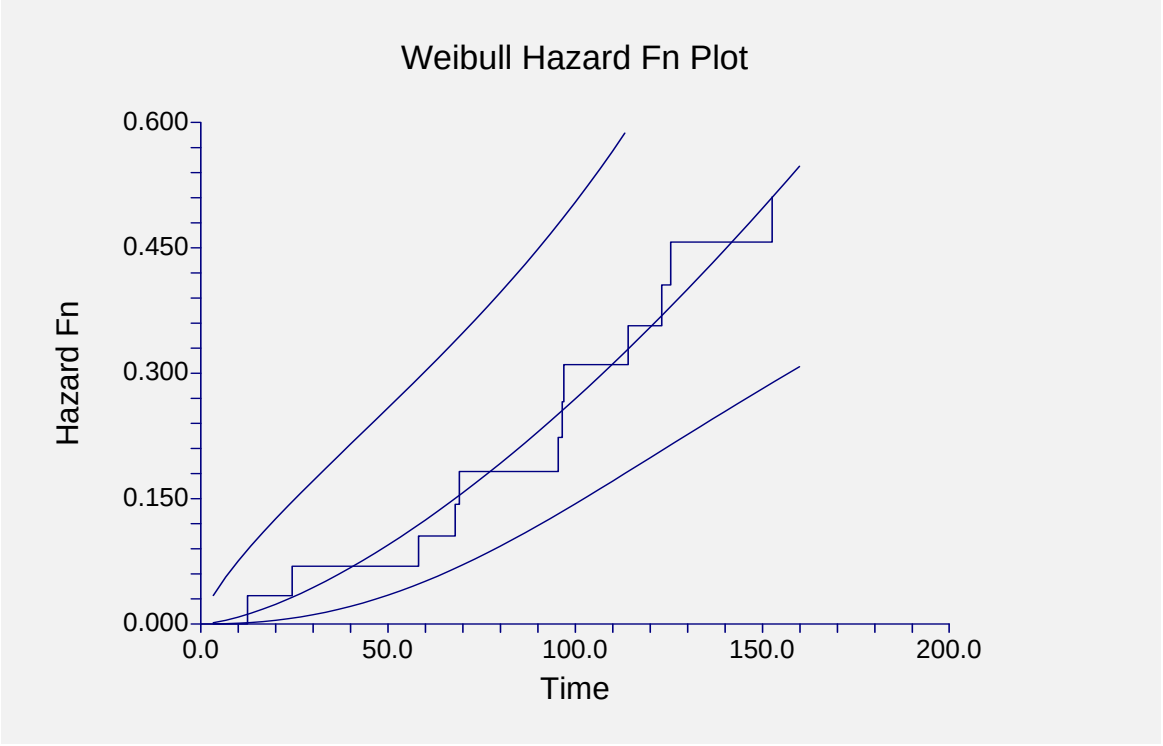
   For the lognormal, exponential, Weibull, and log-logistic distributions, the estimate is given by

   $$T_p = e^{w_p}$$

   For the lognormal base 10 distribution, the estimate is given by

   $$T_p = 10^{w_p}$$

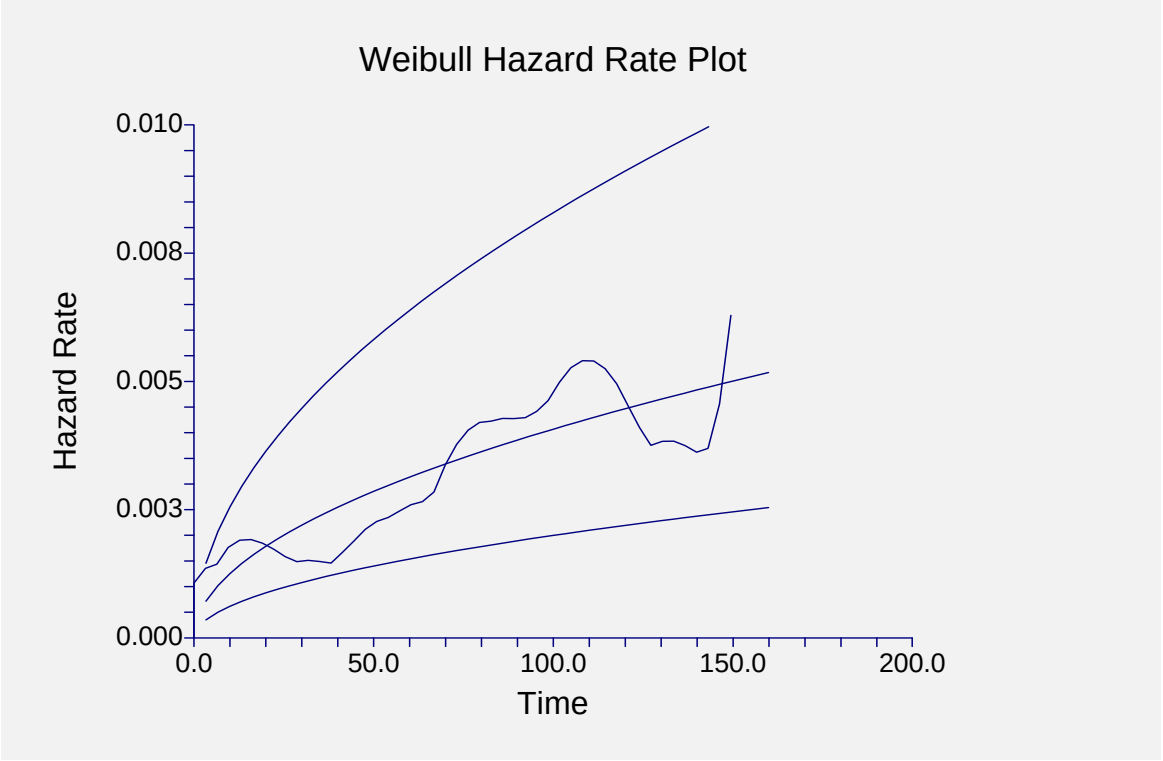
Hazard Function Plot



This plot shows the parametric and nonparametric cumulative hazard functions for the data analyzed. Confidence limits for the parametric cumulative hazard function are also given.

If you have several groups, then a separate line is drawn for each group. The shape of the hazard function is often used to determine an appropriate survival distribution.

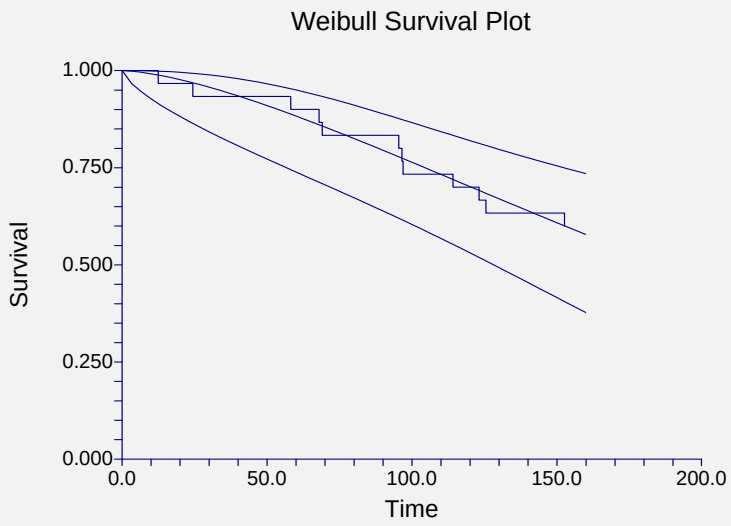
Hazard Rate Plot



This plot shows the parametric and nonparametric hazard rate plots with confidence limits for the parametric hazard rate. This plot is especially useful for studying the shape of the nonparametric hazard rate and comparing that with the parametric hazard rate. When selecting a probability distribution to represent a set of data, it is important to determine if the parametric hazard rate plot has a general shape that is consistent both with the nonparametric hazard rate and with your prior knowledge of the hazard distribution. This plot allows you to make this comparison.

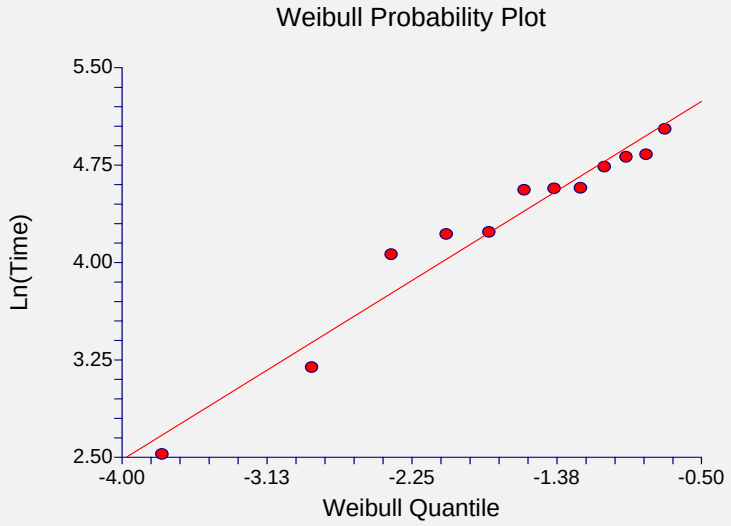
Note that the asymptotic confidence intervals are not well behaved for some distributions. If the confidence intervals seem to have zero width at some point along the plot, you should realize that they fall into this category and ignore them.

Survival Plot



This plot shows the product-limit survivorship function (the step function) as well as the parametric survival plot and associated confidence intervals. If there are several groups, a separate line is drawn for each group.

Probability Plot



This is the Weibull probability plot of these data. The expected quantile of the theoretical distribution is plotted on the horizontal axis. The natural logarithm of the time value is plotted on the vertical axis. Note that censored points are not shown on this plot. Also note that for tied data, only one point is shown for each set of ties.

This plot lets you investigate the goodness of fit of the selected probability distribution to your data. If the points seem to fall along a straight line, the selected probability model may be useful. If the plot shows a downward curve, the value of the threshold parameter, *D*, may need to be

increased. If the plot shows an upward curve, the value of the threshold parameter may need to be decreased. Or you may need to select a different distribution.

You have to decide whether the probability distribution is a good fit to your data by looking at this plot and by comparing the value of the log likelihood to that of other distributions.

Multiple-Censored and Grouped Data

The case of grouped and multiple-censored data cause special problems when creating a probability plot. Remember that the horizontal axis represents the expected quantile from the selected distribution for each (sorted) failure time. In the regular case, we use the rank of the observation in the overall dataset. However, in the case of grouped or multiple-censored data, we use a modified rank. This modified rank, O_j, is computed as follows

$$O_j = O_p + I_j$$

where

$$I_j = \frac{(n+1) - O_p}{1+c}$$

where I_j is the increment for the jth failure; n is the total number of data points, both censored and uncensored; O_p is the order of the previous failure; and c is the number of data points remaining in the data set, including the current data. Implementation details of this procedure may be found in Dodson (1994).

Left censored and interval censored data are treated as failures for making the probability plots.

Example 2 – Distribution Selection

This section presents an example of how to let the program help you pick an appropriate parametric distribution. The data used were shown above and are found in the WEIBULL database.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Distribution (Weibull) Fitting window.

1 Open the WEIBULL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **WEIBULL.S0**.
- Click **Open**.

2 Open the Distribution (Weibull) Fitting window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Distribution (Weibull) Fitting**. The Distribution (Weibull) Fitting procedure will be displayed.
- On the menus, select File, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Distribution (Weibull) Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Time** from the list of variables and then click **Ok**.
- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **Censor** from the list of variables and then click **Ok**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**.
- Select **Find Best** in the **Distribution** box.

4 Set the Derivative constant.

- On the Distribution (Weibull) Fitting window, select the **Options tab**.
- Set **Derivatives** box to **0.00006**.

5 Set the Plots

- On the Distribution (Weibull) Fitting window, select the **Plots tab**.
- Make sure the **Two Plots Per Line** box is checked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum
Failed	12	12	12.5	152.7
Right Censored	1	18	152.7	152.7
Left Censored	0	0		
Interval Censored	0	0		
Total	13	30	12.5	152.7

This report displays a summary of the data that were analyzed. Scan this report to determine if there were any obvious data errors by double-checking the counts and the minimum and maximum.

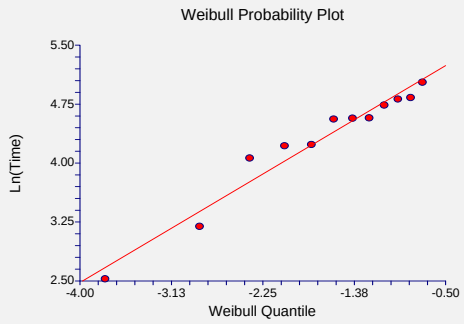
Distribution Fit Summary Section

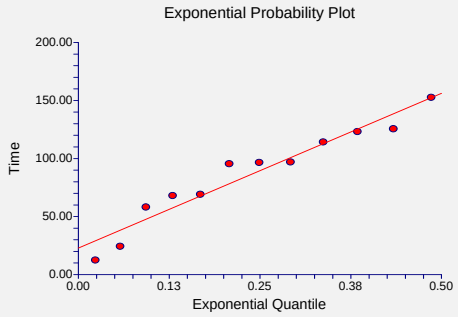
Distribution Fit Summary Section

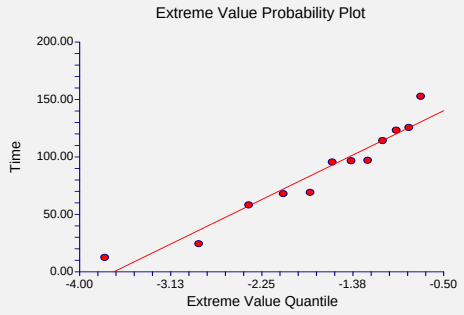
Distribution	Likelihood	Shape	Scale	Threshold
Weibull	-80.05649	1.511543	238.3481	0.0
Loglogistic	-80.11679	5.28008	0.5909371	0.0
Lognormal	-80.38821	5.349999	1.137753	0.0
Lognormal10	-80.38821	2.323475	0.4941201	0.0
Exponential	-81.04864	1	315.4667	0.0
Normal	-81.24539	171.1062	84.88175	0.0
Logistic	-81.74763	169.1118	49.77026	0.0
Extreme Value	-82.1103	189.3399	57.44398	0.0

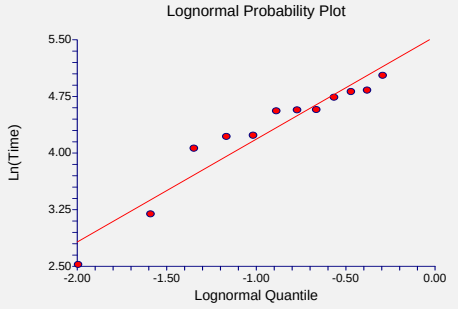
This report displays the values of the log-likelihood for each distribution along with the estimated values of its parameters. Since our desire is to maximize the likelihood, under normal circumstances, we would pick the distribution at the top of the report since it has the largest likelihood value. In this example, we would select the Weibull distribution.

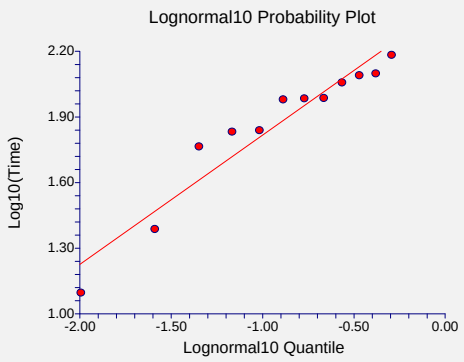
Probability Plots

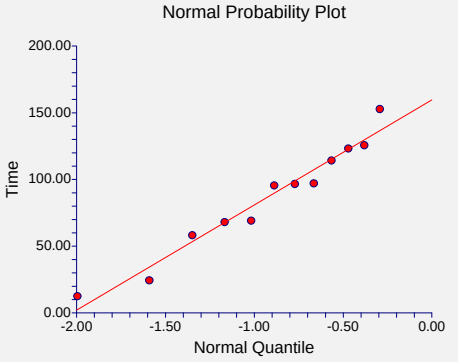


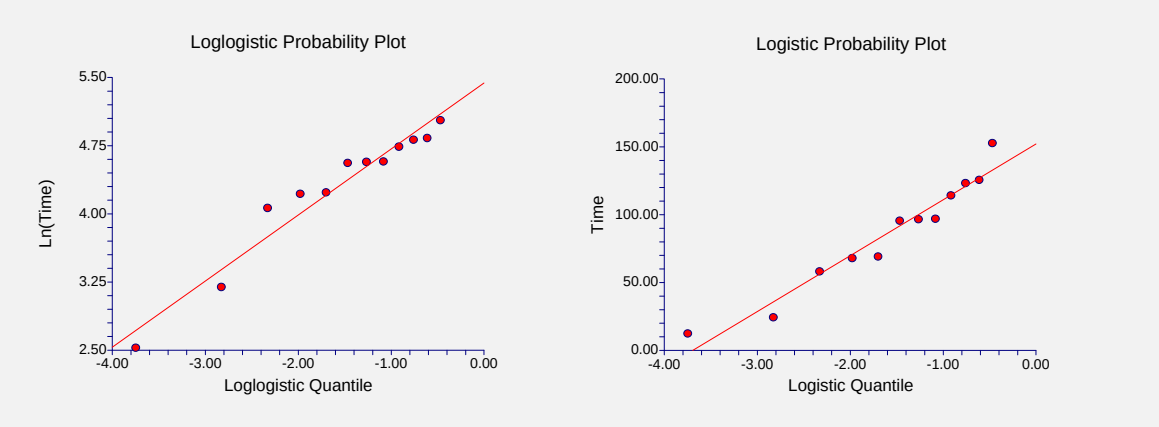












By studying these probability plots, we can determine which distributions fit the data the best. In this example, since there are only a few observations, it is difficult to select one distribution over another. We can see that our candidate from the last section, the Weibull distribution, certainly cannot be removed on the basis of its probability plot. Without further information, our decision would be to select the Weibull distribution to fit these data.

Example 3 – Readout Data

This section presents an example of how to analyze readout data. The data used are found in the READOUT105 database. The table below shows the results of a study to test the failure rate of a particular machine. This study began with 40 test machines. After each time period (24, 72, 168, etc.) the number of machines that had failed since the period began was recorded. This number is entered into the Count variable. Hence, two machines failed during the first 24 hours, one machine failed between 24 and 72 hours, and so on.

After 1500 hours, the study was terminated. Sixteen machines still had not failed. The data are entered in the spreadsheet as shown below.

We have used obvious indicators for censoring. Since the first period begins with a zero time, this entry represents left censored data. We indicate left censoring with an ‘L.’ The next eight rows represent interval censored data. Both beginning and ending times are needed for these entries. We indicate interval censoring with an ‘I.’ The last row corresponds to the sixteen machines that did not fail. These are entered as right censored data, which is indicated with an ‘R.’

READOUT105 dataset

Time1	Time2	Censor	Count
24	0	L	2
72	24	I	1
168	72	I	3
300	168	I	2
500	300	I	2
750	500	I	4
1000	750	I	5
1250	1000	I	1
1500	1250	I	4
1500		R	16

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Distribution (Weibull) Fitting window.

1 Open the READOUT105 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **READOUT105.S0**.
- Click **Open**.

2 Open the Distribution (Weibull) Fitting window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Distribution (Weibull) Fitting**. The Distribution (Weibull) Fitting procedure will be displayed.
- On the menus, select File, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Distribution (Weibull) Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Time1** from the list of variables and then click **Ok**.
- Double-click in the **Start Time Variable** box. This will bring up the variable selection window.
- Select **Time2** from the list of variables and then click **Ok**.
- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **Censor** from the list of variables and then click **Ok**.
- Set **Failed** to **F**.
- Set **Right** to **R**.
- Set **Left** to **L**.
- Set **Interval** to **I**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**.
- Select **Find Best** in the **Distribution** box.

4 Set the Derivative constant.

- On the Distribution Fitting window, select the **Options tab**.
- Set the **Derivatives** box to **0.00006**.

5 Set the Plots

- On the Distribution Fitting window, select the **Plots tab**.
- Make sure the **Two Plots Per Line** box is checked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum
Failed	0	0		
Right Censored	1	16	1500	1500
Left Censored	1	2	24	24
Interval Censored	8	22	24	1500
Total	10	40	24	1500

This report displays a summary of the data that were analyzed. We note that the number of rows and the total count appear to be correct.

Distribution Fit Summary Section

Distribution Fit Summary Section

Distribution	Likelihood	Shape	Scale	Threshold
Weibull	-79.42889	0.8222772	1746.067	0.0
Exponential	-79.96207	1	1631.161	0.0
Loglogistic	-80.27086	7.044066	1.030881	0.0
Lognormal	-81.19075	7.015936	1.886779	0.0
Lognormal10	-81.19075	3.046982	0.8194178	0.0
Normal	-81.44245	1213.697	913.7082	0.0
Logistic	-82.05516	1199.686	563.52	0.0
Extreme Value	-83.09204	1525.271	726.1455	0.0

It appears that the Weibull distribution is a reasonable choice for the parametric distribution, although the shape parameter is less than one. This may point to the need for a nonzero threshold value.

To finish this example, you would view the probability plots. Finally, you would try fitting the Weibull distribution to these data. We will leave that to you to do. Simply change the Distribution box to Weibull and rerun the procedure.

Example 4 – Engine Fan Data

Nelson (1982) gives data on the failure times of seventy diesel engine fans. Twelve of the fans failed during the duration of the test. Fifty-eight of the fans completed the test without failure, so only their running times were recorded. These data are contained in the FANFAILURE database. You can observe the data by opening this database. Note that ‘F’ designates a failure and ‘C’ designates a censored (non-failed) fan.

Two questions were to be answered from these data. First of all, the warranty period for the fan is 8000 hours. Management wanted to know what percentage would fail on or before the warranty period ended. Second, management wanted to know what happens to the failure rate as the fans age.

The following steps will set up the procedure to analyze these data and answer the two questions given above. You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Distribution (Weibull) Fitting window.

1 Open the FANFAILURE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FANFAILURE.S0**.
- Click **Open**.

2 Open the Distribution (Weibull) Fitting window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Distribution (Weibull) Fitting**. The Distribution (Weibull) Fitting procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Distribution (Weibull) Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Hours** from the list of variables and then click **Ok**.
- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **Censor** from the list of variables and then click **Ok**.
- Set Failed to **F**.
- Set Right to **C**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**.
- Select **Weibull** in the Distribution box.

4 Set the Reports

- On the Distribution (Weibull) Fitting window, select the **Reports tab**.
- Set the Times box to **1000:15000(1000)**.
- Set the Probability Decimal Places box to **7**. This forces the display of seven decimal places.

5 Set the Hazard Rate Plot

- On the Distribution (Weibull) Fitting window, select the **Haz Rt Plot tab**.
- Click on the **Vertical Axis Tick Label Settings** button. This will allow you to reset the reference numbers on the vertical (Y) axis.
- Set Decimals to **7**.
- Set Max Characters to **10**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

We will show only those portions of the printout that are necessary to answer the two questions that were posed at the beginning of this section.

Parameter Estimation Section

Weibull Parameter Estimation Section

Parameter	Probability Plot Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Shape	1.202295	1.058446	0.2683138	0.6440074	1.739588
Scale	17283.81	26296.85	12254.1	10549.97	65547.48
Threshold	0	0			
Log Likelihood		-135.1527			
Mean	16250.13	25715.61			
Median	12742.28	18600.24			
Mode	3925.094	1703.919			
Sigma	13574.96	24306.58			

This report shows the estimated parameters. We are particularly interested to see that the shape parameter is almost exactly one. The confidence limits for the estimated shape parameter include one between them. Remember that when the shape parameter is one, the Weibull distribution reduces to the exponential distribution, a distribution which 'has no memory.' From this, we get an indication that the failure pattern of the fans does not change over time. That is, the failure rate does not change as the fans get older.

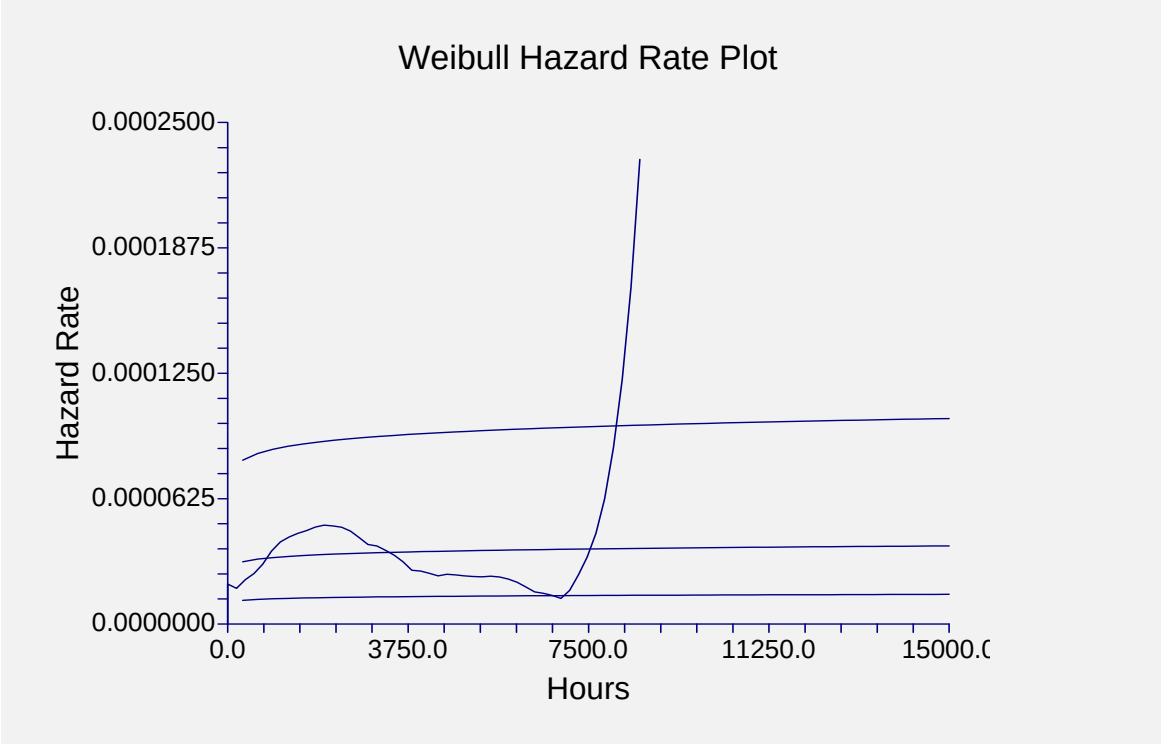
Parametric Failure Distribution Section

Weibull Failure Distribution Section

Failure Time	Prob Plot Estimate of Failure	Max Like Estimate of Failure	95% Lower Conf. Limit of Failure	95% Upper Conf. Limit of Failure
1000.0	0.0319854	0.0309247	0.0104713	0.0894825
2000.0	0.0720737	0.0633292	0.0288586	0.1359862
3000.0	0.1146709	0.0956044	0.0501483	0.1782088
4000.0	0.1581267	0.1273811	0.0718910	0.2203039
5000.0	0.2015573	0.1584891	0.0927279	0.2635977
6000.0	0.2444119	0.1888367	0.1120577	0.3082572
7000.0	0.2863235	0.2183722	0.1297484	0.3538902
8000.0	**0.3270409**	**0.2470672**	**0.1458995**	**0.3998867**
9000.0	0.3663914	0.2749082	0.1606924	0.4456080
10000.0	0.4042590	0.3018915	0.1743176	0.4904755
11000.0	0.4405694	0.3280202	0.1869456	0.5340032
12000.0	0.4752803	0.3533026	0.1987208	0.5758038
13000.0	0.5083738	0.3777505	0.2097617	0.6155832
14000.0	0.5398509	0.4013782	0.2201653	0.6531311
15000.0	0.5697275	0.4242020	0.2300113	0.6883101

This report presents the estimated failure proportions at various time periods. We note that at 8000 hours, the maximum likelihood estimate for the proportion failing is 0.247. The 95% confidence limits are 0.146 to 0.400. That is, almost 25% of the fans can be expected to fail by 8000 hours—a very high failure rate. Management will have to change the fans to decrease the proportion failing!

Hazard Rate Plot



This plot shows both the parametric and nonparametric estimates of the hazard rates. First, we analyze the nonparametric estimate. Notice that the line wanders up and then down, but it does not extend outside the confidence limits of the parametric hazard rate. The sharp rise at the end of the plot is due to a lack of data in this region and should be ignored. We see that the parametric estimate of the hazard rate, the middle horizontal line, is a reasonable approximation for the nonparametric line. The above considerations again lead us to conclude that the failure rates do not change with age.

This ends this example. Notice how quickly we have been able to answer the two questions posed by management. The only task that we did not complete was to make sure that the Weibull distribution was appropriate for these data. A quick look at the probability plot will show you that it is.

Chapter 551

Beta Distribution Fitting

Introduction

This module fits the beta probability distributions to a complete set of individual or grouped data values. It outputs various statistics and graphs that are useful in reliability and survival analysis.

The beta distribution is useful for fitting data which have an absolute maximum (and minimum). It finds some application as a lifetime distribution.

Technical Details

The four-parameter beta distribution is indexed by two shape parameters (P and Q) and two parameters representing the minimum (A) and maximum (B). We will not estimate A and B, but rather assume that they are known parameters.

Using these symbols, the beta density function may be written as

$$f(t|P,Q,A,B) = \frac{1}{B(P,Q)} \frac{(t-A)^{P-1}(B-t)^{Q-1}}{(B-A)^{P+Q-1}}, \quad P>0,\ Q>0,\ A<t<B$$

where

$$B(P,Q) = \frac{\Gamma(P)\Gamma(Q)}{\Gamma(P+Q)}$$

Making the transformation

$$X = \frac{(t-A)}{(B-A)}$$

results in the two-parameter beta distribution. This is also known as the standardized form of the beta distribution. In this case, the density function is

$$f(x|P,Q) = \frac{1}{B(P,Q)} x^{P-1}(1-x)^{Q-1}, \quad P>0,\ Q>0,\ 0<x<1$$

Reliability Function

The reliability (or survivorship) function, *R(t)*, gives the probability of surviving beyond time *t*. For the beta distribution, the reliability function is

$$R(T) = 1 - \int_{A}^{T} f\left(t|P,Q,A,B\right)dt$$

where the integral is known as the *incomplete beta function ratio*.

The conditional reliability function, *R(t,T)*, may also be of interest. The is the reliability of an item given that it has not failed by time *T*. The formula for the conditional reliability is

$$R(t) = \frac{R(T+t)}{R(T)}$$

Hazard Function

The hazard function represents the instantaneous failure rate. For this distribution, the hazard function is

$$h(t) = \frac{f(t)}{R(t)}$$

Kaplan-Meier Product-Limit Estimator

The production limit estimator is covered in the Distribution Fitting chapter and will not be repeated here.

Data Structure

Beta datasets require only a failure time variable. Censored data may not be fit with this program. An optional count variable which gives the number of items occurring at that time period. If the count variable is omitted, all counts are assumed to be one.

The table below shows the results of a study to test failure rate of a particular machine which has a maximum life of 100 hours. This particular experiment began with 10 items under test. After all items had failed, the experiment was stopped. These data are contained on the BETA database.

BETA dataset (subset)

Time
23.5
50.1
65.3
68.9
70.4
77.3
81.6

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Variable

Time Variable

This variable contains the failure times. Note that negative time values and time values less than the minimum parameter are treated as missing values. Zero time values are replaced by the value in the Zero Time Replacement option.

These time values represent elapsed times. If your data has dates (such as the failure date), you must subtract the starting date so that you can analyze the elapsed time.

Zero Time Replacement

Under normal conditions, a respondent beginning the study is “alive” and cannot “die” until after some small period of time has elapsed. Hence, a time value of zero is not defined and is ignored (treated as a missing value). If a zero time value does occur on the database, it is replaced by this positive amount. If you do not want zero time values replaced, enter a “0.0” here.

This option would be used when a “zero” on the database does not actually mean zero time. Instead, it means that the response occurred before the first reading was made and so the actual survival time is only known to be less.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Frequency Variable

Frequency Variable

This variable gives the number of individuals (the count or frequency) at a given failure (or censor) time. When omitted, each row receives a frequency of one. Frequency values should be positive integers.

Options

Product Limit and Hazard Conf. Limits Method

The standard nonparametric estimator of the reliability function is the Product Limit estimator. This option controls the method used to estimate the confidence limits of the estimated reliability. The options are Linear, Log Hazard, Arcsine Square Root, and Nelson-Aalen. The formulas used by these options were presented in the Technical Details section of the Distribution Fitting chapter. Although the Linear (Greenwood) is the most commonly used, recent studies have

shown either the Log Hazard or the Arsine Square Root Hazard are better in the sense that they require a smaller sample size to be accurate.

Beta Minimum

This option sets the value of the minimum. Usually, this value is zero. All data values used must be greater than this value.

Beta Maximum

This option sets the value of the maximum. Often, this value is one. All data values used must be less than this value.

Options – Search

Maximum Iterations

Many of the parameter estimation algorithms are iterative. This option assigns a maximum to the number of iterations used in any one algorithm. We suggest a value of about 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained.

Minimum Relative Change

This value is used to control the iterative algorithms used in parameter estimation. When the relative change in any of the parameters is less than this amount, the iterative procedure is terminated.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report - Percentiles Report

These options indicate whether to display the corresponding report.

Select Plots

Survivorship Plot - Probability Plot

These options indicate whether to display the corresponding report or plot.

Report Options

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run

into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Survival and Haz Rt Calculation Values

Percentiles

This option specifies a list of percentiles (range 1 to 99) at which the reliability (survivorship) is reported. The values should be separated by commas.

Specify sequences with a colon, putting the increment inside parentheses after the maximum in the sequence. For example: 5:25(5) means 5,10,15,20,25 and 1:5(2),10:20(2) means 1,3,5,10,12,14,16,18,20.

Times

This option specifies a list of times at which the percent surviving is reported. Individual values are separated by commas. You can specify a sequence by specifying the minimum and maximum separate by a colon and putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers. Use ‘(10)’ alone to specify ten values between zero and the maximum value found in the data.

Report Options – Decimal Places

Time Decimals

This option specifies the number of decimal places shown on reported time values.

Plot Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the name of the group variable.

Survival and Cum Haz Plot Tabs

These options control the attributes of the survival curves and the hazard curves.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Beta Plot Tab

These options control the attributes of the beta reliability curve.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Beta Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Time Values

Specify the number of points along the curve at which calculations are made. This controls the resolution of the curve. Usually, values between 50 and 200 produce good results.

Beta Plot Settings – Plot Contents

Product-Limit Curve

Indicate whether to overlay the product-limit curve on this plot. Overlaying the PL curve lets you determine whether the beta curve is a reasonable approximation to this curve.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

These options control the attributes of the beta probability plot.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Y Scaling

Indicate whether the vertical scaling on all means plots should uniform across all plots.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Prob Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Let us emphasize that this probability plot uses the Scatter Plot style files, not the Probability Plot style files.

Plotting Position - F(T)

The probability plot shows time on the vertical axis and the gamma quantile on the horizontal axis. This option specifies the method used to determine F(t) which used to calculate the vertical

plotting positions on the probability plot. Note that method selected here also influences the probability plot estimates of the parameters.

The five alternatives available are

- **Median (j-0.3)/(n+0.4)**

  The most popular method is to calculate the median rank for each sorted data value. That is, this is the value for the j^{th} sorted time value. Since the median rank requires extensive calculations, this approximation to the median rank is often used.

$$F(t_j) = \frac{j - 0.3}{n + 0.4}$$

- **Median (Exact)**

  The most popular method is to calculate the median rank for each data value. This is the median rank of the j^{th} sorted time value out of n values. The exact value of the median rank is calculated using the formula

$$F(t_j) = \frac{1}{1 + \frac{n - j + 1}{j} F_{0.50;\, 2(n-j+1);\, 2j}}$$

- **Mean j/(n+1)**

  The mean rank is sometimes recommended. In this case, the formula is

$$F(t_j) = \frac{j}{n + 1}$$

- **White (j-3/8)/(n+1/4)**

  A formula proposed by White is sometimes recommended. The formula is

$$F(t_j) = \frac{j + 3/8}{n + 1/4}$$

- **(j-0.5)/n**

  The following formula is sometimes used

$$F(t_j) = \frac{j - 0.5}{n}$$

Prob Plot Settings – Plot Contents

Trend Line

This option controls whether the trend (least squares) line is calculated and displayed.

Residuals from Trend Line

This option controls whether the vertical deviations from the trend line are displayed. Displaying these residuals may let you see departures from linearity more easily.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines & Symbols Tab

These options specify the attributes of the lines used for each group in the hazard curves and survival curves and the symbols used for each group in the probability plots.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Plotting Symbols

Symbol 1 - 15

These options specify the symbols used in the plot of each group. The first symbol is used by the first group, the second symbol by the second group, and so on. These symbols are provided to allow the various groups to be easily identified, even on black and white printers.

Clicking on a symbol box (or the small button to the right of the symbol box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Beta Distribution

This section presents an example of how to fit a beta distribution. The data used were shown above and are found in the BETA database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Beta Distribution Fitting window.

1 Open the BETA dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **BETA.S0**.
- Click **Open**.

2 Open the Beta Distribution Fitting window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Beta Distribution Fitting**. The Beta Distribution Fitting procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Beta Distribution Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Time** from the list of variables and then click **Ok**.
- Click in the **Beta Maximum** box. Enter **100** for the maximum value.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum	Average	Sigma
Failed	10	10	23.5	95.3	70.8	21.2021

This report displays a summary of the data that were analyzed. Scan this report to determine if there were any obvious data errors by double-checking the counts and the minimum and maximum.

Parameter Estimation Section

Parameter Estimation Section

Parameter	Method of Moments Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Minimum (A)	0	0			
Maximum (B)	100	100			
P	2.548055	3.301583	1.485834	0.3894027	6.213764
Q	1.050893	1.414615	0.577846	0.2820573	2.547172
Log Likelihood		-3.403845			
Mean	70.8	70.00519			
Median	74.91825	73.002			
Mode	96.81711	84.73547			
Sigma	21.2021	19.16614			

This report displays parameter estimates along with standard errors and confidence limits in the maximum likelihood case.

Method of Moments Estimate

By equating the theoretical moments with the data moments, the following estimates are obtained.

$$\tilde{P} = \frac{\left[\frac{m_1 - A}{B - A}\right]^2 \left[1 - \frac{m_1 - A}{B - A}\right]}{\left[\frac{m_2}{(B - A)^2}\right]} - \left[\frac{m_1 - A}{B - A}\right]$$

$$\tilde{Q} = \frac{\left[\frac{m_1 - A}{B - A}\right] \left[1 - \frac{m_1 - A}{B - A}\right]}{\left[\frac{m_2}{(B - A)^2}\right]} - \tilde{P}$$

where m_1 is the usual estimator of the mean and m_2 is the usual estimate of the standard deviation.

Maximum Likelihood Estimates of A, C, and D

These estimates maximize the likelihood function. The maximum likelihood equations are

$$\psi\left(\hat{P}\right) - \psi\left(\hat{P} + \hat{Q}\right) = \frac{1}{n}\sum_{j=1}^{n} \log\left(\frac{t_j - A}{B - A}\right)$$

$$\psi\left(\hat{Q}\right) - \psi\left(\hat{P} + \hat{Q}\right) = \frac{1}{n}\sum_{j=1}^{n} \log\left(\frac{B - t_j}{B - A}\right)$$

where $\psi(x)$ is the digamma function.

The formulas for the standard errors and confidence limits come from the inverse of the Fisher information matrix, *{f(i,j)}*. The standard errors are given as the square roots of the diagonal elements *f(1,1)* and *f(2,2)*. The confidence limits for *P* are

$$\hat{P}_{lower,1-\alpha/2} = \hat{P} - z_{1-\alpha/2}\sqrt{f(1,1)}$$

$$\hat{P}_{upper,1-\alpha/2} = \hat{P} + z_{1-\alpha/2}\sqrt{f(1,1)}$$

The confidence limits for *Q* are

$$\hat{Q}_{lower,1-\alpha/2} = \hat{Q} - z_{1-\alpha/2}\sqrt{f(2,2)}$$

$$\hat{Q}_{upper,1-\alpha/2} = \hat{Q} + z_{1-\alpha/2}\sqrt{f(2,2)}$$

Log Likelihood

This is the value of the log likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Mean

This is the mean time to failure (MTTF). It is the mean of the random variable (failure time) being studied given that the beta distribution provides a reasonable approximation to your data's actual distribution.

The formula for the mean is

$$Mean = A + \frac{P(B-A)}{P+Q}$$

Median

The median of the beta distribution is the value of *t* where *F(t)=0.5*.

$$Median = A + I(0.5, P, Q)$$

where *I(0.5,P,C)* is the incomplete beta function.

Mode

The mode of the beta distribution is given by

$$Mode = A + \frac{(P-1)(B-A)}{P+Q-2}$$

when $A > 1$ and *D* otherwise.

Sigma

This is the standard deviation of the failure time. The formula for the standard deviation (sigma) of a beta random variable is

$$\sigma = \sqrt{\frac{PQ(B-A)^2}{(P+Q)^2(P+Q+1)}}$$

Inverse of Fisher Information Matrix

Inverse of Fisher Information Matrix

Parameter	Scale	Shape
Scale	2.207702	0.6725335
Shape	0.6725335	0.333906

This table gives the inverse of the Fisher information matrix for the two-parameter beta. These values are used in creating the standard errors and confidence limits of the parameters and reliability statistics. The approximate Fisher information matrix is given by the 2-by-2 matrix whose elements are

$$f(1,1) = \frac{\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)}{n\left(\psi'\left(\hat{P}\right)\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)\left\{\psi'\left(\hat{P}\right)+\psi'\left(\hat{Q}\right)\right\}\right)}$$

$$f(1,2) = f(2,1) = \frac{\psi'\left(\hat{P}+\hat{Q}\right)}{n\left(\psi'\left(\hat{P}\right)\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)\left\{\psi'\left(\hat{P}\right)+\psi'\left(\hat{Q}\right)\right\}\right)}$$

$$f(2,2) = \frac{\psi'\left(\hat{P}\right) - \psi'\left(\hat{P}+\hat{Q}\right)}{n\left(\psi'\left(\hat{P}\right)\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)\left\{\psi'\left(\hat{P}\right)+\psi'\left(\hat{Q}\right)\right\}\right)}$$

where $\psi'(z)$ is the trigamma function and *n* represents the sample size.

Kaplan-Meier Product-Limit Survival Distribution

Kaplan-Meier Product-Limit Survival Distribution

Failure Time	Lower 95% C.L. Survival	Estimated Survival	Upper 95% C.L. Survival	Lower 95% C.L. Hazard	Estimated Hazard	Upper 95% C.L. Hazard	Sample Size
23.5	0.714061	0.900000	1.000000	0.000000	0.105361	0.336786	10
50.1	0.552082	0.800000	1.000000	0.000000	0.223144	0.594059	9
65.3	0.415974	0.700000	0.984026	0.016103	0.356675	0.877132	8
68.9	0.296364	0.600000	0.903636	0.101328	0.510826	1.216168	7
70.4	0.190102	0.500000	0.809898	0.210848	0.693147	1.660192	6
77.3	0.096364	0.400000	0.703636	0.351494	0.916291	2.339626	5
81.6	0.015974	0.300000	0.584026	0.537810	1.203973	4.136778	4
85.7	0.000000	0.200000	0.447918	0.803145	1.609438		3
89.9	0.000000	0.100000	0.285939	1.251978	2.302585		2
95.3							1

Confidence Limits Method: Linear (Greenwood)

This report displays the Kaplan-Meier product-limit survival distribution and hazard function along with confidence limits. The formulas used were presented in the Technical Details section earlier in this chapter. Note that these estimates do not use the beta distribution in any way. They are the nonparametric estimates and are completely independent of the distribution that is being fit. We include them for reference.

Note that the Sample Size is given for each time period. As time progresses, participants are removed from the study, reducing the sample size. Hence, the survival results near the end of the study are based on only a few participants and are therefore less reliable. This shows up in a widening of the confidence limits.

Reliability Section

Reliability Section

Fail Time	ProbPlot Estimated Reliability	MLE Estimated Reliability
5.0	0.999474	0.999900
10.0	0.996931	0.999030
15.0	0.991393	0.996366
20.0	0.982123	0.990777
25.0	0.968503	0.981102
30.0	0.949995	0.966195
35.0	0.926119	0.944961
40.0	0.896446	0.916390
45.0	0.860585	0.879593
50.0	0.818187	0.833836
55.0	0.768940	0.778582
60.0	0.712568	0.713545
65.0	0.648840	0.638750
70.0	0.577573	0.554623
75.0	0.498648	0.462119
80.0	0.412033	0.362916
85.0	0.317839	0.259761
90.0	0.216436	0.157159
95.0	0.108834	0.063202
100.0	0.000000	0.000000

This report displays the estimated reliability (survivorship) at the time values that were specified in the Times option of the Reports Tab. Reliability may be thought of as the probability that failure occurs after the given failure time. Thus, (using the ML estimates) the probability is 0.944961 that failure will not occur until after 35 hours.

Two reliability estimates are provided. The first uses the method of moments estimates and the second uses the maximum likelihood estimates. Confidence limits are not available. The formulas used are as follows.

Estimated Reliability

The reliability (survivorship) is calculated using the beta distribution as

$$\hat{R}(t) = \hat{S}(t) = 1 - I\left(\frac{t - A}{B - A}; P, Q\right)$$

Percentile Section

Percentile Section

Percentile	MOM Failure Time	MLE Failure Time
5.00	30.0	33.9
10.00	39.5	42.4
15.00	46.3	48.3
20.00	51.9	53.2
25.00	56.8	57.3
30.00	61.0	61.0
35.00	64.9	64.3
40.00	68.5	67.4
45.00	71.8	70.3
50.00	74.9	73.0
55.00	77.9	75.6
60.00	80.7	78.2
65.00	83.3	80.6
70.00	85.9	83.1
75.00	88.4	85.5
80.00	90.8	87.9
85.00	93.1	90.4
90.00	95.4	92.9
95.00	97.7	95.8

This report displays failure time percentiles using the method of moments and the maximum likelihood estimates. No confidence limit formulas are available.

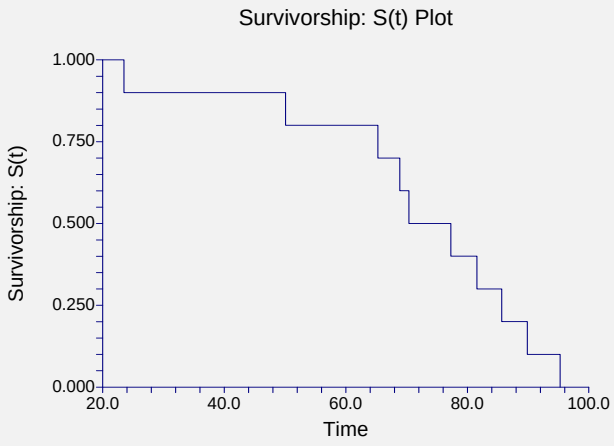
The formulas used are

Estimated Percentile

The time percentile at P (which ranges between zero and one hundred) is calculated using

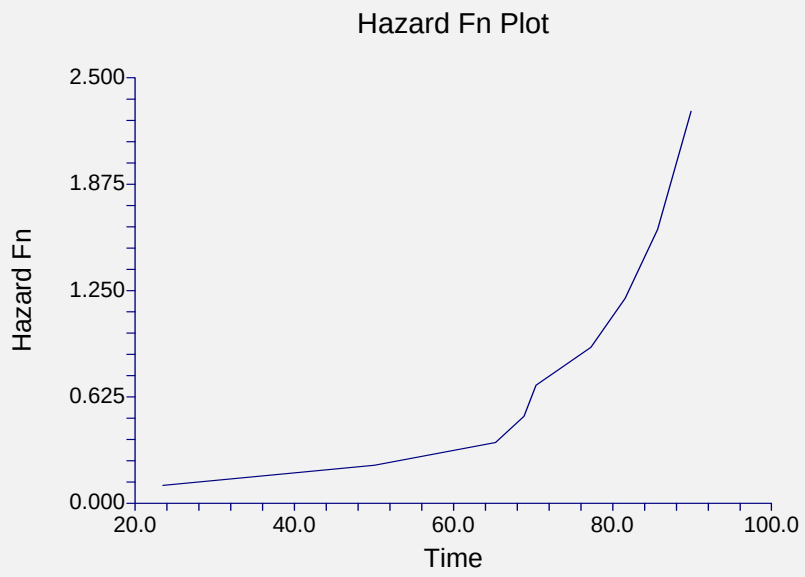
$$\hat{t}_p = \left[A + I(p; A, C)\left(B - A\right)\right] \times 100$$

Product-Limit Survivorship Plot



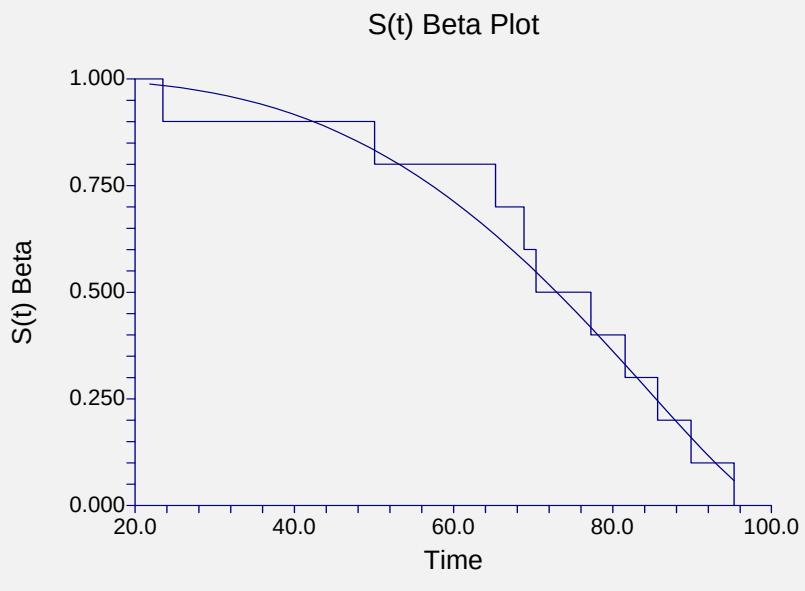
This plot shows the product-limit survivorship function for the data analyzed. If you have several groups, a separate line is drawn for each group. The step nature of the plot reflects the nonparametric product-limit survival curve.

Hazard Function Plot



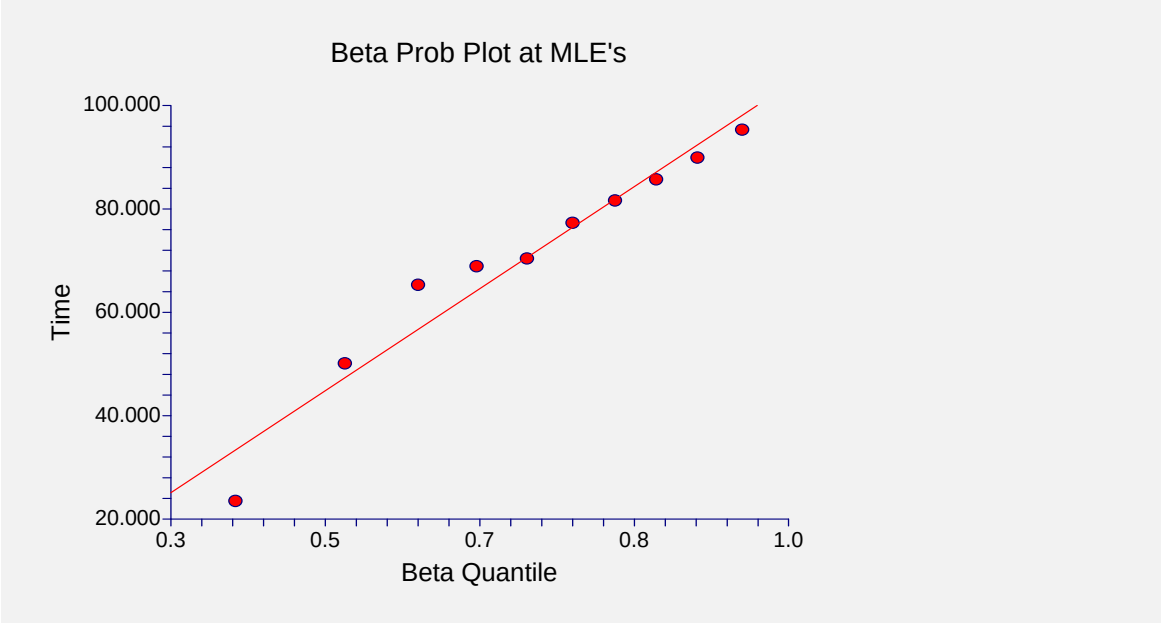
This plot shows the cumulative hazard function for the data analyzed. If you have several groups, then a separate line is drawn for each group. The shape of the hazard function is often used to determine an appropriate survival distribution.

Beta Reliability Plot



This plot shows the product-limit survival function (the step function) and the beta distribution overlaid. If you have several groups, a separate line is drawn for each group.

Beta Probability Plots



This is a beta probability plot for these data. The expected quantile of the theoretical distribution is plotted on the horizontal axis. The time value is plotted on the vertical axis. Also note that for grouped data, only one point is shown for each group.

This plot lets you investigate the goodness of fit of the beta distribution to your data. If the points seem to fall along a straight line, the beta probability model may be useful. You have to decide whether the beta distribution is a good fit to your data by looking at this plot and by comparing the value of the log likelihood to that of other distributions.

Grouped Data

The case of grouped data causes special problems when creating a probability plot. Remember that the horizontal axis represents the expected quantile from the beta distribution for each (sorted) failure time. In the regular case, we used the rank of the observation in the overall dataset. However, in case of grouped data, we most use a modified rank. This modified rank, O_j, is computed as follows

$$O_j = O_p + I_j$$

where

$$I_j = \frac{(n+1) - O_p}{1+c}$$

were I_j is the increment for the jth failure; n is the total number of data points; O_p is the order of the previous failure; and c is the number of data points remaining in the data set, including the current data. Implementation details of this procedure may be found in Dodson (1994).

Chapter 552

Gamma Distribution Fitting

Introduction

This module fits the gamma probability distributions to a complete or censored set of individual or grouped data values. It outputs various statistics and graphs that are useful in reliability and survival analysis.

The gamma distribution competes with the Weibull distribution as a model for lifetime. Since it is more complicated to deal with mathematically, it has been used less. While the Weibull is a purely heuristic model (approximating the data well), the gamma distribution does arise as a physical model since the sum of exponential random variables results in a gamma random variable.

At times, you may find that the distribution of log lifetime follows the gamma distribution.

The Three-Parameter Gamma Distribution

The three-parameter gamma distribution is indexed by a shape , a scale, and a threshold parameter. Many symbols have been used to represent these parameters in the statistical literature. We have selected the symbols *A*, *C*, and *D* for the shape, scale, and threshold. Our choice of symbols was made to make remembering their meanings easier. That is, just remember shApe, sCale, and thresholD and you will remember the general meaning of each symbol. Using these symbols, the three parameter gamma density function may be written as

$$f(t|A,C,D) = \frac{1}{C\Gamma(A)}\left(\frac{t-D}{C}\right)^{A-1} e^{-\frac{t-D}{C}}, \quad A>0, C>0, -\infty<D<\infty, t>D$$

Shape Parameter - A

This parameter controls the shape of the distribution. When $A = 1$, the gamma distribution is identical to the exponential distribution. When $C = 2$ and $A = v/2$, where v is an integer, the gamma becomes the chi-square distribution with v degrees of freedom. When A is restricted to integers, the gamma distribution is referred to as the Erlang distribution used in queueing theory.

Scale Parameter - C

This parameter controls the scale of the data. When *C* becomes large, the gamma distribution approaches the normal distribution.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. When *D* is set to zero, we obtain the two parameter gamma distribution. In *NCSS*, the threshold is not an estimated quantity but rather a fixed constant. Care should be used in using the threshold parameter because it forces the probability of failure to be zero between 0 and *D*.

Reliability Function

The reliability (or survivorship) function, *R(t)*, gives the probability of surviving beyond time *t*. For the gamma distribution, the reliability function is

$$R(t) = 1 - I(t)$$

where *I(t)* in this case represents the incomplete gamma function.

The conditional reliability function, *R(t,T)*, may also be of interest. The is the reliability of an item given that it has not failed by time *T*. The formula for the conditional reliability is

$$R(t) = \frac{R(T+t)}{R(T)}$$

Hazard Function

The hazard function represents the instantaneous failure rate. For this distribution, the hazard function is

$$h(t) = \frac{f(t)}{R(t)}$$

Kaplan-Meier Product-Limit Estimator

The production limit estimator is covered in the Distribution Fitting chapter and will not be repeated here.

Data Structure

Most gamma datasets require two (and often three) variables: the failure time variable, an optional censor variable formed by entering a zero for a censored observation or a one for a failed observation, and an optional count variable which gives the number of items occurring at that time period. If the censor variable is omitted, all time values represent observations from failed items. If the count variable is omitted, all counts are assumed to be one.

The table below shows the results of a study to test failure rate of a particular machine. This particular experiment began with 30 items under test. After the twelfth item failed at 152.7 hours, the experiment was stopped. The remaining eighteen observations were censored. That is, we know that they will fail at some time in the future. These data are contained on the WEIBULL database.

WEIBULL dataset

Time	Censor	Count
12.5	1	1
24.4	1	1
58.2	1	1
68.0	1	1
69.1	1	1
95.5	1	1
96.6	1	1
97.0	1	1
114.2	1	1
123.2	1	1
125.6	1	1
152.7	1	1
152.7	0	18

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Variable

Time Variable

This variable contains the failure times. Note that negative time values and time values less than the threshold parameter are treated as missing values. Zero time values are replaced by the value in the Zero Time Replacement option.

These time values represent elapsed times. If your data has dates (such as the failure date), you must subtract the starting date so that you can analyze the elapsed time.

Zero Time Replacement

Under normal conditions, a respondent beginning the study is “alive” and cannot “die” until after some small period of time has elapsed. Hence, a time value of zero is not defined and is ignored (treated as a missing value). If a zero time value does occur on the database, it is replaced by this positive amount. If you do not want zero time values replaced, enter a “0.0” here.

This option would be used when a “zero” on the database does not actually mean zero time. Instead, it means that the response occurred before the first reading was made and so the actual survival time is only known to be less.

Censor Variable

Censor Variable

This optional variable contains the censor indicator variable. The value is set to zero for censored observations and one for failed observations.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Frequency Variable

Frequency Variable

This variable gives the number of individuals (the count or frequency) at a given failure (or censor) time. When omitted, each row receives a frequency of one. Frequency values should be positive integers.

Options

Threshold Value

This option controls the setting of the threshold parameter. When this value is set to zero (which is the default) the two-parameter gamma distribution is fit. You can put in a fixed, nonzero value for *D* here.

A cautionary note is needed. The maximum value that *D* can have is the minimum time value. If the minimum time is a censored observation you may be artificially constraining *D* to an inappropriately low value. It may make more sense to ignore these censored observations or to fit the two-parameter gamma.

Product Limit and Hazard Conf. Limits Method

The standard nonparametric estimator of the reliability function is the Product Limit estimator. This option controls the method used to estimate the confidence limits of the estimated reliability. The options are Linear, Log Hazard, Arcsine Square Root, and Nelson-Aalen. The formulas used by these options were presented in the Technical Details section of the Distribution Fitting chapter. Although the Linear (Greenwood) is the most commonly used, recent studies have shown either the Log Hazard or the Arsine Square Root Hazard are better in the sense that they require a smaller sample size to be accurate.

Options – Probability Plot

Least Squares Model

When a probability plot is used to estimate the parameters of the gamma model, this option designates which variable (time or frequency) is used as the dependent variable.

- **F=A+B(Time)**

  On the probability plot, *F* is regressed on Time and the resulting intercept and slope are used to estimate the gamma parameters. See the discussion of probability plots below for more information.

- **Time=A+B(F)**

  On the probability plot, Time is regressed on *F* and the resulting intercept and slope are used to estimate the gamma parameters.

Shape Values

This options specifies values for the shape parameter, *A*, at which probability plots are to be generated. You can use a list of numbers separated by blanks or commas. Or, you can use the special list format: e.g. 0.5:2.0(0.5) which means 0.5 1.0 1.5 2.0. All values must be greater than zero.

Final Shape Value

This option specifies the shape parameter value that is used in the reports. The value in the list above that is closest to this value is used.

Use of this option usually requires two runs. In the first run, the probability plots of several trial *A* values are considered. The value of *A* for which the probability plot appears the straightest (in which all points fall along an imaginary straight line) is determined and used in a second run. Or, you may decide to use a value near the maximum likelihood estimate of *A*.

Options – Search

Maximum Iterations

Many of the parameter estimation algorithms are iterative. This option assigns a maximum to the number of iterations used in any one algorithm. We suggest a value of about 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained.

Minimum Relative Change

This value is used to control the iterative algorithms used in parameter estimation. When the relative change in any of the parameters is less than this amount, the iterative procedure is terminated.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report - Percentiles Report

These options indicate whether to display the corresponding report.

Select Plots

Survivorship Plot - Probability Plot

These options indicate whether to display the corresponding report or plot.

Report Options

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Survival and Haz Rt Calculation Values

Percentiles

This option specifies a list of percentiles (range 1 to 99) at which the reliability (survivorship) is reported. The values should be separated by commas.

Specify sequences with a colon, putting the increment inside parentheses after the maximum in the sequence. For example: 5:25(5) means 5,10,15,20,25 and 1:5(2),10:20(2) means 1,3,5,10,12,14,16,18,20.

Times

This option specifies a list of times at which the percent surviving is reported. Individual values are separated by commas. You can specify a sequence by specifying the minimum and maximum

separate by a colon and putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers. Use '(10)' alone to specify ten values between zero and the maximum value found in the data.

Time Decimals

This option specifies the number of decimal places shown on reported time values.

Plot Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the name of the group variable.

Survival and Cum Haz Plot Tabs

These options control the attributes of the survival curves and the hazard curves.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Y Scaling

Indicate whether the vertical scaling on all means plots should uniform across all plots.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Gamma Plot Tab

These options control the attributes of the gamma reliability curve.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Y Scaling

Indicate whether the vertical scaling on all means plots should uniform across all plots.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Gamma Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Time Values

Specify the number of points along the curve at which calculations are made. This controls the resolution of the curve. Usually, values between 50 and 200 produce good results.

Gamma Plot Settings – Plot Contents

Product-Limit Curve

Indicate whether to overlay the product-limit curve on this plot. Overlaying the PL curve lets you determine whether the gamma curve is a reasonable approximation to this curve.

Confidence Limits

Indicate whether to display the confidence limits of the reliability curve on the plot.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

These options control the attributes of the gamma probability plot. Remember that the probability plot can only be generated for a given value of the shape parameter. This value is set in the Final Shape option of the Search tab, which was described earlier.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Y Scaling

Indicate whether the vertical scaling on all means plots should uniform across all plots.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Prob Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Let us emphasize that this probability plot uses the Scatter Plot style files, not the Probability Plot style files.

Plotting Position - F(T)

The probability plot shows time on the vertical axis and the gamma quantile on the horizontal axis. This option specifies the method used to determine F(t) which used to calculate the vertical plotting positions on the probability plot. Note that method selected here also influences the probability plot estimates of the parameters.

The five alternatives available are

- **Median (j-0.3)/(n+0.4)**
  The most popular method is to calculate the median rank for each sorted data value. That is, this is the value for the j^{th} sorted time value. Since the median rank requires extensive calculations, this approximation to the median rank is often used.
  $$F(t_j) = \frac{j - 0.3}{n + 0.4}$$

- **Median (Exact)**
  The most popular method is to calculate the median rank for each data value. This is the median rank of the j^{th} sorted time value out of n values. The exact value of the median rank is calculated using the formula
  $$F(t_j) = \frac{1}{1 + \frac{n - j + 1}{j} F_{0.50;\, 2(n-j+1);\, 2j}}$$

- **Mean j/(n+1)**
  The mean rank is sometimes recommended. In this case, the formula is
  $$F(t_j) = \frac{j}{n + 1}$$

- **White (j-3/8)/(n+1/4)**
  A formula proposed by White is sometimes recommended. The formula is

$$F(t_j) = \frac{j + 3/8}{n + 1/4}$$

- **(j-0.5)/n**
  The following formula is sometimes used

$$F(t_j) = \frac{j - 0.5}{n}$$

Prob Plot Settings – Plot Contents

Trend Line

This option controls whether the trend (least squares) line is calculated and displayed.

Residuals from Trend Line

This option controls whether the vertical deviations from the trend line are displayed. Displaying these residuals may let you see departures from linearity more easily.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines & Symbols Tab

These options specify the attributes of the lines used for each group in the hazard curves and survival curves and the symbols used for each group in the probability plots.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Plotting Symbols

Symbol 1 - 15

These options specify the symbols used in the plot of each group. The first symbol is used by the first group, the second symbol by the second group, and so on. These symbols are provided to allow the various groups to be easily identified, even on black and white printers.

Clicking on a symbol box (or the small button to the right of the symbol box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Gamma Distribution

This section presents an example of how to fit a gamma distribution. The data used were shown above and are found in the WEIBULL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Gamma Distribution Fitting window.

1 Open the WEIBULL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **WEIBULL.S0**.
- Click **Open**.

2 Open the Gamma Distribution Fitting window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Gamma Distribution Fitting**. The Gamma Distribution Fitting procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Gamma Distribution Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Time** from the list of variables and then click **Ok**.

- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **Censor** from the list of variables and then click **Ok**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**.

4 Specify the plots.

- On the Gamma Distribution Fitting window, select the **Plots tab**.
- Check the **Confidence Limits** box.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum	Average	Sigma
Failed	12	12	12.5	152.7	86.41666	41.66633
Censored	1	18	152.7	152.7		
Total	13	30	12.5	152.7		

Type of Censoring: Singly

This report displays a summary of the data that were analyzed. Scan this report to determine if there were any obvious data errors by double-checking the counts and the minimum and maximum.

Parameter Estimation Section

Parameter Estimation Section

Parameter	Probability Plot Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Shape	2	2.407362	0.9228407	0.598627	4.216096
Scale	107.21	85.21822	36.31201	36.96837	196.4422
Threshold	0	0			
Log Likelihood		-80.6078			
Mean	214.42	205.1511			
Median	179.9356	177.551			
Mode	107.21	119.9329			
Sigma	151.6178	132.2218			

This report displays parameter estimates along with standard errors and confidence limits in the maximum likelihood case. In this example, we have set the threshold parameter to zero so we are fitting the two-parameter gamma distribution.

Probability Plot Estimate

This estimation procedure uses the data from the gamma probability plot to estimate the parameters. The estimation formula depends on which option was selected for the Least Squares Model. Note that the value of A is given—only C is estimated from the plot.

Least Squares Model: F=A+B(Time)

Using simple linear regression through the origin, we obtain the estimate of *C* as

$$\tilde{C} = \text{slope}$$

Least Squares Model: Time=A+B(F)

Using simple linear regression through the origin, we obtain the estimate of *C* as

$$\tilde{C} = \frac{1}{slope}$$

Maximum Likelihood Estimates of A, C, and D

These estimates maximize the likelihood function. The formulas for the standard errors and confidence limits come from the inverse of the Fisher information matrix, *{f(i,j)}*. The standard errors are given as the square roots of the diagonal elements *f(1,1)* and *f(2,2)*. The confidence limits for *A* are

$$\hat{A}_{lower,1-\alpha/2} = \hat{A} - z_{1-\alpha/2}\sqrt{f(1,1)}$$

$$\hat{A}_{upper,1-\alpha/2} = \hat{A} + z_{1-\alpha/2}\sqrt{f(1,1)}$$

The confidence limits for *C* are

$$\hat{C}_{lower,1-\alpha/2} = \frac{\hat{C}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{f(2,2)}}{\hat{C}}\right\}}$$

$$\hat{C}_{upper,1-\alpha/2} = \hat{C}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{f(2,2)}}{\hat{C}}\right\}$$

Log Likelihood

This is the value of the log likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Mean

This is the mean time to failure (MTTF). It is the mean of the random variable (failure time) being studied given that the gamma distribution provides a reasonable approximation to your data's actual distribution.

The formula for the mean is

$$Mean = D + AC$$

Median

The median of the gamma distribution is the value of *t* where *F(t)=0.5*.

$$Median = D + I(0.5, A, C)$$

where *I(0.5,A,C)* is the incomplete gamma function.

Mode

The mode of the gamma distribution is given by

$$Mode = D + C(A-1)$$

when $A > 1$ and D otherwise.

Sigma

This is the standard deviation of the failure time. The formula for the standard deviation (sigma) of a gamma random variable is

$$\sigma = C\sqrt{A}$$

Inverse of Fisher Information Matrix

Inverse of Fisher Information Matrix

Parameter	Shape	Scale
Shape	0.8516349	-30.14704
Scale	-30.14704	1318.562

This table gives the inverse of the Fisher information matrix for the two-parameter gamma. These values are used in creating the standard errors and confidence limits of the parameters and reliability statistics. These statistics are very difficult to calculate directly for the gamma distribution when censored data are present. We use a large sample approximation that has been suggested by some authors. These results are only accurate when the shape parameter is greater than two.

The approximate Fisher information matrix is given by the 2-by-2 matrix whose elements are

$$f(1,1) = \frac{\hat{A}}{n\left(\hat{A}\psi'\left(\hat{A}\right)-1\right)}$$

$$f(1,2) = f(2,1) = \frac{-\hat{C}}{n\left(\hat{A}\psi'\left(\hat{A}\right)-1\right)}$$

$$f(2,2) = \frac{\hat{C}^2\psi'\left(\hat{A}\right)}{n\left(A\psi'\left(\hat{A}\right)-1\right)}$$

where $\psi'(z)$ is the trigamma function and n represents the number of failed items (does not include censored items).

Kaplan-Meier Product-Limit Survival Distribution

Kaplan-Meier Product-Limit Survival Distribution

Failure Time	Lower 95% C.L. Survival	Estimated Survival	Upper 95% C.L. Survival	Lower 95% C.L. Hazard	Estimated Hazard	Upper 95% C.L. Hazard	Sample Size
12.5	0.902433	0.966667	1.000000	0.000000	0.033902	0.102661	30
24.4	0.844073	0.933333	1.000000	0.000000	0.068993	0.169517	29
58.2	0.792648	0.900000	1.000000	0.000000	0.105361	0.232376	28
68.0	0.745025	0.866667	0.988308	0.011760	0.143101	0.294338	27
69.1	0.699975	0.833333	0.966692	0.033875	0.182322	0.356711	26
95.5	0.656864	0.800000	0.943136	0.058545	0.223144	0.420278	25
96.6	0.615318	0.766667	0.918016	0.085541	0.265703	0.485616	24
97.0	0.575091	0.733333	0.891576	0.114765	0.310155	0.553227	23
114.2	0.536018	0.700000	0.863982	0.146203	0.356675	0.623588	22
123.2	0.497980	0.666667	0.835354	0.179900	0.405465	0.697196	21
125.6	0.460893	0.633333	0.805774	0.215952	0.456758	0.774590	20
152.7	0.424695	0.600000	0.775305	0.254499	0.510826	0.856383	19
152.7+							18

Confidence Limits Method: Linear (Greenwood)

This report displays the Kaplan-Meier product-limit survival distribution and hazard function along with confidence limits. The formulas used were presented in the Technical Details section earlier in this chapter. Note that these estimates do not use the gamma distribution in any way. They are the nonparametric estimates and are completely independent of the distribution that is being fit. We include them for reference.

Note that censored observations are marked with a plus sign on their time value. The survival and hazard functions are not calculated for censored observations.

Also note that the Sample Size is given for each time period. As time progresses, participants are removed from the study, reducing the sample size. Hence, the survival results near the end of the study are based on only a few participants and are therefore less reliable. This shows up in a widening of the confidence limits.

Reliability Section

Reliability Section

Fail Time	ProbPlot Estimated Reliability	MLE Estimated Reliability	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
8.0	0.997351	0.998953	0.929195	1.000000
16.0	0.989912	0.994798	0.921475	1.000000
24.0	0.978387	0.987065	0.910223	1.000000
32.0	0.963401	0.975768	0.895440	1.000000
40.0	0.945512	0.961121	0.877317	1.000000
48.0	0.925214	0.943434	0.856131	1.000000
56.0	0.902947	0.923062	0.832187	1.000000
64.0	0.879099	0.900376	0.805796	0.994957
72.0	0.854012	0.875743	0.777251	0.974235
80.0	0.827987	0.849515	0.746827	0.952203
88.0	0.801290	0.822024	0.714773	0.929274
96.0	0.774151	0.793576	0.681318	0.905834
104.0	0.746772	0.764451	0.646672	0.882229
112.0	0.719328	0.734901	0.611036	0.858766
120.0	0.691969	0.705152	0.574602	0.835701
128.0	0.664826	0.675402	0.537563	0.813240
136.0	0.638009	0.645826	0.500114	0.791538
144.0	0.611611	0.616574	0.462450	0.770698
152.0	0.585711	0.587777	0.424773	0.750781
160.0	0.560373	0.559543	0.387282	0.731804

This report displays the estimated reliability (survivorship) at the time values that were specified in the Times option of the Reports Tab. Reliability may be thought of as the probability that failure occurs after the given failure time. Thus, (using the ML estimates) the probability is 0.975768 that failure will not occur until after 32 hours. The 95% confidence for this estimated probability is 0.895440 to 1.000000.

Two reliability estimates are provided. The first uses the parameters estimated from the probability plot and the second uses the maximum likelihood estimates. Confidence limits are calculated for the maximum likelihood estimates. The formulas used are as follows.

Estimated Reliability

The reliability (survivorship) is calculated using the gamma distribution as

$$\hat{R}(t) = \hat{S}(t) = 1 - I\left(t - D; A, C\right)$$

Confidence Limits for Reliability

The confidence limits for this estimate are computed using the following formulas. Note that these estimates lack accuracy when A is less than 2.0.

$$\hat{R}_{upper}(t) = \hat{R}(t) - z_{1-\alpha/2}\sqrt{Var\left(\hat{R}(t)\right)}$$

$$\hat{R}_{upper}(t) = \hat{R}(t) + z_{1-\alpha/2}\sqrt{Var\left(\hat{R}(t)\right)}$$

where

$$Var\left(\hat{R}(t)\right) \cong \frac{\phi^2\left(\hat{\beta}\right)}{n}\left[\frac{2\left(t-D\right)^2}{\hat{C}\hat{A}^2} - \left(2\hat{C}-1\right)\left(1+\frac{\hat{\beta}}{2\sqrt{\hat{C}}}\right)\left(1+\frac{3\hat{\beta}}{2\sqrt{\hat{C}}}\right)\right]$$

where $\phi\left(z\right)$ is the standard normal density and

$$\hat{\beta} = \frac{\left(t-D\right)/\hat{A} - \hat{C}}{\sqrt{\hat{C}}}$$

Percentile Section

Percentile Section

Percentile	MLE Failure Time
5.00	45.2
10.00	64.1
15.00	79.9
20.00	94.2
25.00	107.9
30.00	121.4
35.00	134.9
40.00	148.6
45.00	162.7
50.00	177.6
55.00	193.2
60.00	210.1
65.00	228.5
70.00	249.1
75.00	272.6
80.00	300.4
85.00	335.1
90.00	382.2
95.00	459.4

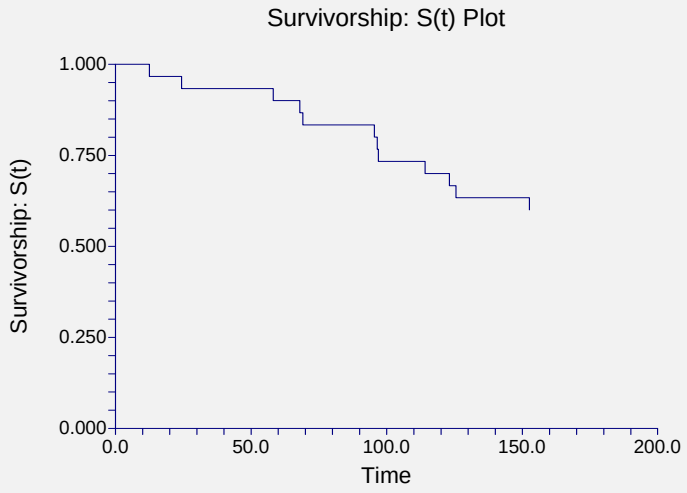
This report displays failure time percentiles using the maximum likelihood estimates. No confidence limit formulas are available.

Estimated Percentile

The time percentile at P (which ranges between 0 and 100) is calculated using

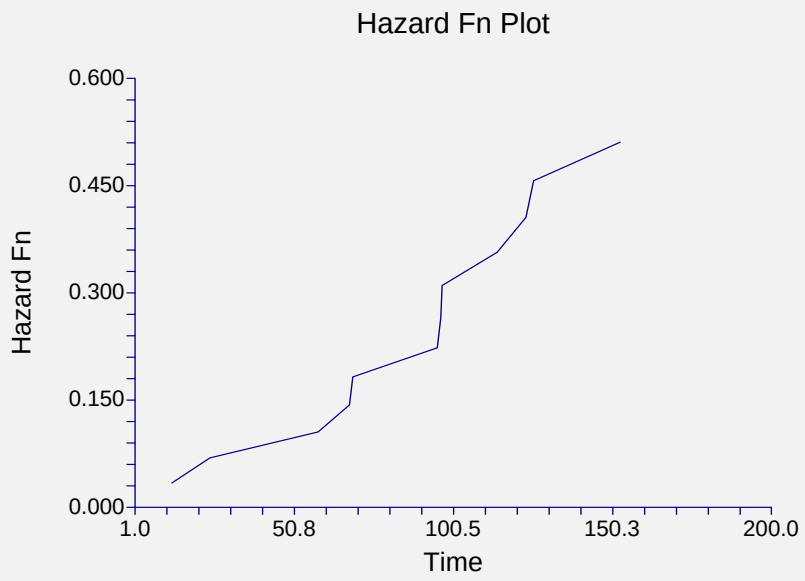
$$\hat{t}_p = \left[D + I(p; A, C)\right] \times 100$$

Product-Limit Survivorship Plot



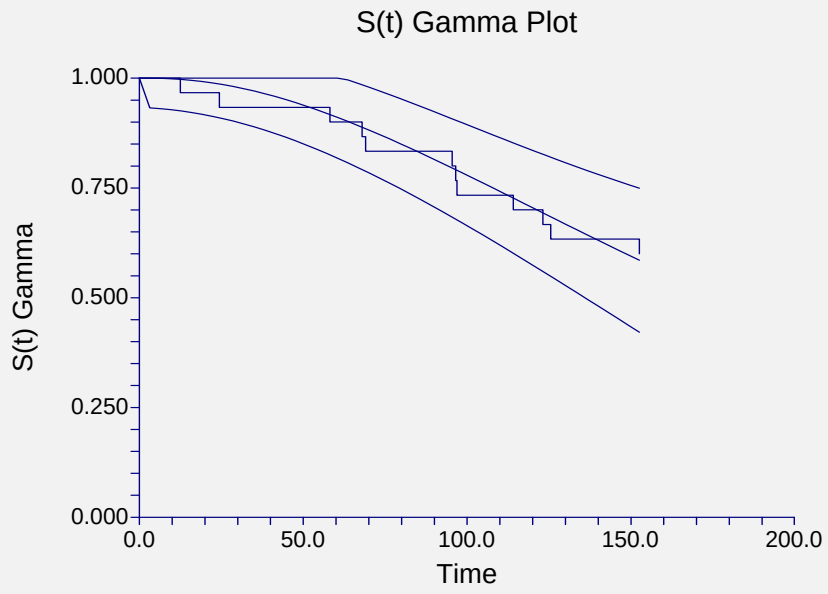
This plot shows the product-limit survivorship function for the data analyzed. If you have several groups, a separate line is drawn for each group. The step nature of the plot reflects the nonparametric product-limit survival curve.

Hazard Function Plot



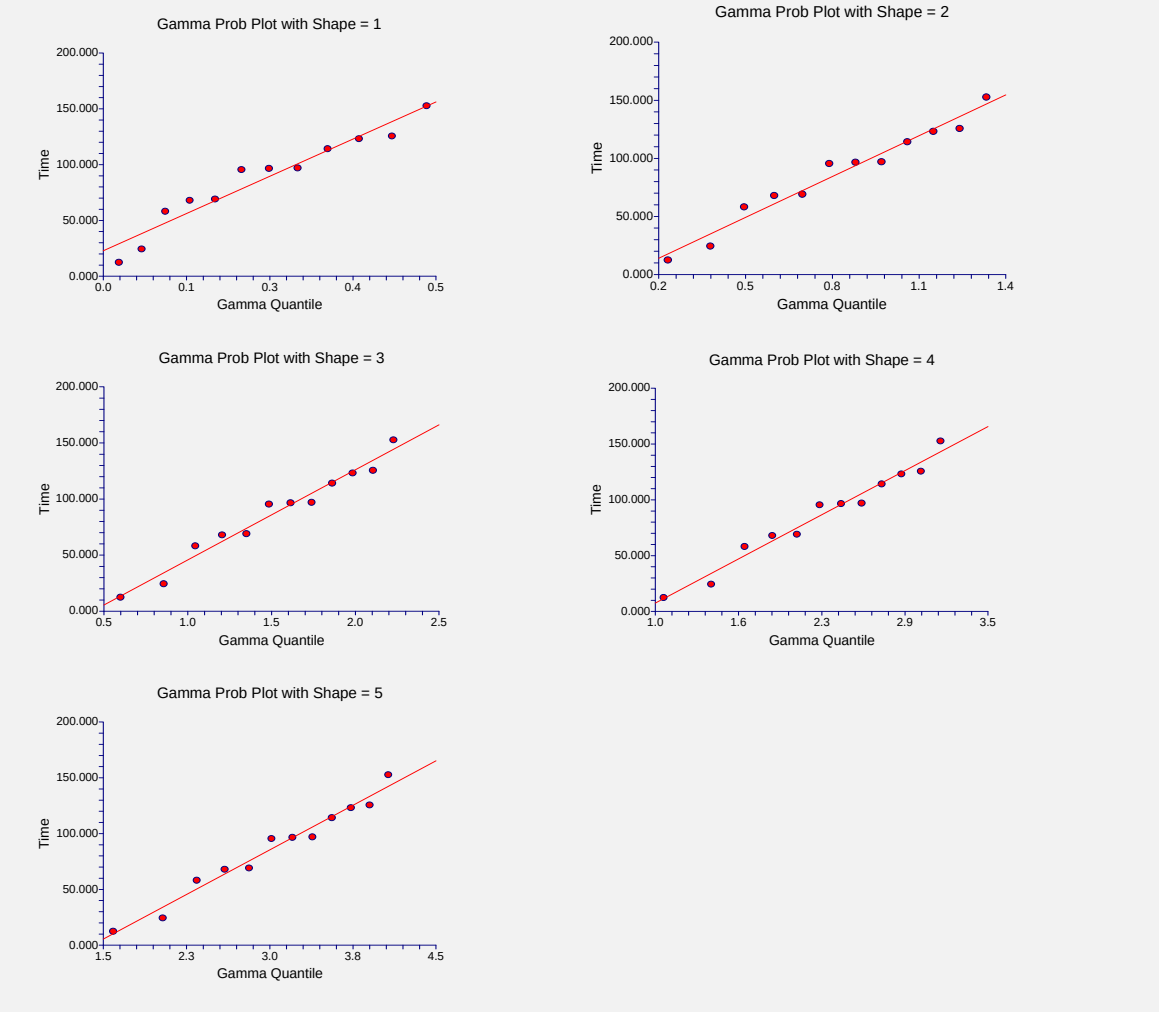
This plot shows the cumulative hazard function for the data analyzed. If you have several groups, then a separate line is drawn for each group. The shape of the hazard function is often used to determine an appropriate survival distribution.

Gamma Reliability Plot



This plot shows the product-limit survival function (the step function) and the gamma distribution overlaid. The confidence limits are also displayed. If you have several groups, a separate line is drawn for each group.

Gamma Probability Plots



There is a gamma probability plot for each specified value of the shape parameter (set in the Prob.Plot Shape Values option of the Search tab). The expected quantile of the theoretical distribution is plotted on the horizontal axis. The time value is plotted on the vertical axis. Note that censored points are not shown on this plot. Also note that for grouped data, only one point is shown for each group.

These plot let determine an appropriate value of *A*. They also let you investigate the goodness of fit of the gamma distribution to your data. You have to decide whether the gamma distribution is a good fit to your data by looking at these plots and by comparing the value of the log likelihood to that of other distributions.

For this particular set of data, it appears that *A* equal two or three would work just fine. Note that the maximum likelihood estimate of *A* is 2.4—right in between!

Multiple-Censored and Grouped Data

The case of grouped, or multiple-censored, data cause special problems when creating a probability plot. Remember that the horizontal axis represents the expected quantile from the gamma distribution for each (sorted) failure time. In the regular case, we used the rank of the observation in the overall dataset. However, in case of grouped or multiple-censored data, we most use a modified rank. This modified rank, O_j, is computed as follows

$$O_j = O_p + I_j$$

where

$$I_j = \frac{(n+1) - O_p}{1+c}$$

were I_j is the increment for the jth failure; n is the total number of data points, both censored and uncensored; O_p is the order of the previous failure; and c is the number of data points remaining in the data set, including the current data. Implementation details of this procedure may be found in Dodson (1994).

Chapter 555

Kaplan-Meier Curves (Logrank Tests)

Introduction

This procedure computes the nonparametric Kaplan-Meier and Nelson-Aalen estimates of survival and associated hazard rates. It can fit complete, right censored, left censored, interval censored (readout), and grouped data values. It outputs various statistics and graphs that are useful in reliability and survival analysis.

It also performs several logrank tests and provides both the parametric and randomization test significance levels.

Overview of Survival Analysis

We will give a brief introduction to the subject in this section. For a complete account of survival analysis, we suggest the book by Klein and Moeschberger (1997).

Survival analysis is the study of the distribution of life times. That is, it is the study of the elapsed time between an initiating event (birth, start of treatment, diagnosis, or start of operation) and a terminal event (death, relapse, cure, or machine failure). The data values are a mixture of complete (terminal event occurred) and censored (terminal event has not occurred) observations. From the data values, the survival analyst makes statements about the survival distribution of the failure times. This distribution allows questions about such quantities as survivability, expected life time, and mean time to failure to be answered.

Let *T* be the elapsed time until the occurrence of a specified event. The event may be death, occurrence of a disease, disappearance of a disease, appearance of a tumor, etc. The probability distribution of *T* may be specified using one of the following basic functions. Once one of these functions has been specified, the others may be derived using the mathematical relationships presented.

1. Probability density function, *f(t)*. This is the probability that an event occurs at time *t*.

2. Cumulative distribution function, *F(t)*. This is the probability that an individual survives until time *t.*

$$F(t) = \int_0^t f(x)dx$$

3. Survival function, *S(T)*. This is the probability that an individual survives beyond time *T*. This is usually the primary quantity of interest. It is estimated using the nonparametric Kaplan-Meier curve.

$$\begin{aligned} S(T) &= \int_T^\infty f(x)dx \\ &= 1 - F(T) \end{aligned}$$

$$\begin{aligned} S(T) &= \exp\left[-\int_0^T h(x)dx \right] \\ &= \exp\left[-H(T) \right] \end{aligned}$$

4. Hazard rate, *h(T)*. This is the probability that an individual at time *T* experiences the event in the next instant. It is a fundamental quantity in survival analysis. It is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, and the inverse of Mill's ratio in economics. The empirical hazard rate may be used to identify the appropriate probability distribution of a particular mechanism, since each distribution has a different hazard rate function. Some distributions have a hazard rate that decreases with time, others have a hazard rate that increases with time, some are constant, and some exhibit all three behaviors at different points in time.

$$h(T) = \frac{f(T)}{S(T)}$$

5. Cumulative hazard function, *H*(*T*). This is integral of *h*(*T*) from 0 to *T*.

$$\begin{aligned} H(T) &= \int_0^T h(x)dx \\ &= -\ln\left[S(T) \right] \end{aligned}$$

Nonparametric Estimators of Hazard and Survival

All of the following results are from Klein and Moeschberger (1997).

The recommended nonparametric estimator of the survival distribution, *S*(*T*), is the Kaplan-Meier product-limit estimator. The recommended nonparametric estimator of the cumulative hazard function, *H*(*T*), is the Nelson-Aalen estimator. Although each of these estimators could be used to estimate the other quantity using the relationship

$$H(T) = -\ln\left[S(T) \right]$$

or

$$S(T) = \exp\left[-H(T)\right]$$

this is not recommended.

The following notation will be used to define both of these estimators. Let $t = 1, \cdots, M$ index the *M* unique termination (failure or death) times $T_1, T_2, ..., T_M$. Note that *M* does not include duplicate times or times at which only censored observations occur. Associated with each of these failure times is an entry time E_t at which the subject began to be observed. Usually, these entry times are taken to be zero. If positive entry times are specified, the data are said to have been *left truncated*. When data are left truncated, it is often necessary to define a minimum time, *A*, below which failures are not considered. When a positive *A* is used, the unconditional survival function *S*(*T*) is changed to a conditional survival function *S*(*T*|*T*>*A*).

The set of all failures (deaths) that occur at time T_t is referred to as D_t and the number in this set is given by d_t. The *risk set* at *t*, R_t, is the set of all individuals that are at risk immediately before time T_t. This set includes all individuals whose entry and termination times include T_t. That is, R_t is made up of all individuals with times such that $E_j < T_t \le T_j$ and $A \le T_t$. The number of individuals in the risk set is given by r_t.

Kaplan-Meier Product-Limit Estimator

Using the above notation, the Kaplan-Meier product-limit estimator is defined as follows in the range of time values for which there are data.

$$\hat{S}(T) = \begin{cases} 1 & \text{if } T_{\min} > T \\ \prod_{A \le T_i \le T} \left[1 - \dfrac{d_i}{r_i}\right] & \text{if } T_{\min} \le T \end{cases}$$

The variance of *S(T)* is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(T)\right] = \hat{S}(T)^2 \sum_{A \le T_i \le T} \frac{d_i}{r_i(r_i - d_i)}$$

Pointwise Confidence Intervals of Survival

A pointwise confidence interval for the survival probability at a specific time T_0 of $S(T_0)$ is represented by two confidence limits which have been constructed so that the probability that the true survival probability lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire survival function lies within the band. When these are plotted with the survival curve, these limits must be interpreted on an individual, point by point, basis.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended.

Linear (Greenwood) Pointwise Confidence Interval for S(T)

This estimator may be used to create a confidence interval at a specific time point T_0 of $S(T_0)$ using the formula

$$\hat{S}(T_0) \pm z_{1-\alpha/2}\sigma_S(T_0)$$

where

$$\sigma_S^2(T_0) = \frac{\hat{V}\left[\hat{S}(T_0)\right]}{\hat{S}^2(T_0)}$$

and *z* is the appropriate value from the standard normal distribution.

Log-Transformed Pointwise Confidence Interval for S(T)

Better confidence limits may be calculated using the logarithmic transformation of the hazard functions. These limits are

$$\hat{S}(T_0)^{1/\theta}, \hat{S}(T_0)^{\theta}$$

where

$$\theta = \exp\left\{\frac{z_{1-\alpha/2}\sigma_S(T_0)}{\log[\hat{S}(T_0)]}\right\}$$

ArcSine-Square Root Pointwise Confidence Interval for S(T)

Another set of confidence limits using an improving transformation is given by the formula

$$\sin^2\left\{\max\left[0, \arcsin\left\{\hat{S}(T_0)\right\}^{1/2} - 0.5z_{1-\alpha/2}\sigma_S(T_0)\left(\frac{\hat{S}(T_0)}{1-\hat{S}(T_0)}\right)^{1/2}\right]\right\}$$
$$\le S(T_0) \le$$
$$\sin^2\left\{\min\left[\frac{\pi}{2}, \arcsin\left\{\hat{S}(T_0)\right\}^{1/2} + 0.5z_{1-\alpha/2}\sigma_S(T_0)\left(\frac{\hat{S}(T_0)}{1-\hat{S}(T_0)}\right)^{1/2}\right]\right\}$$

Nelson-Aalen Hazard Estimator

The Nelson-Aalen estimator is recommended as the best estimator of the cumulative hazard function, *H*(*T*). This estimator is give as

$$\tilde{H}(T) = \begin{cases} 0 & if\ T_{\min} > T \\ \sum_{A \le T_i \le T} \frac{d_i}{r_i} & if\ T_{\min} \le T \end{cases}$$

Three estimators of the variance of this estimate are mentioned on page 34 of Therneau and Grambsch (2000). These estimators differ in the way they model tied event times. When there are no event time ties, they give almost identical results.

1. Simple (Poisson) Variance Estimate

This estimate assumes that event time ties occur because of rounding and a lack of measurement precision. This estimate is the largest of the three, so it gives the widest, most conservative, confidence limits. The formula for this estimator, derived assuming a Poisson model for the number of deaths, is

$$\sigma^2_{\tilde{H}1}(T) = \sum_{A \le T_i \le T} \frac{d_i}{r_i^2}$$

2. Plug-in Variance Estimate

This estimate also assumes that event time ties occur because of rounding and a lack of measurement precision. The formula for this estimator, derived by substituting sample quantities in the theoretical variance formula, is

$$\sigma^2_{\tilde{H}2}(T) = \sum_{A \le T_i \le T} \frac{d_i\left(r_i - d_i\right)}{r_i^3}$$

Note that when $r_i = 1$, a '1' is substituted for $\left(r_i - d_i\right) / r_i$ in this formula.

3. Binomial Variance Estimate

This estimate assumes that event time ties occur because the process is fundamentally discrete rather than due to lack of precision and/or rounding. The formula for this estimator, derived assuming a binomial model for the number of events, is

$$\sigma^2_{\tilde{H}3}(T) = \sum_{A \le T_i \le T} \frac{d_i\left(r_i - d_i\right)}{r_i^2\left(r_i - 1\right)}$$

Note that when $r_i = 1$, a '1' is substituted for $\left(r_i - d_i\right) / \left(r_i - 1\right)$ in this formula.

Which Variance Estimate to Use

Therneau and Grambsch (2000) indicate that, as of the writing of their book, there is no clear-cut champion. The simple estimate is often suggested because it is always largest and thus gives the widest, most conservative confidence, confidence limits. In practice, there is little difference between them and the choice of which to use will make little difference in the final interpretation of the data. We have included all three since each occurs alone in various treatises on survival analysis.

Pointwise Confidence Intervals of Cumulative Hazard

A pointwise confidence interval for the cumulative hazard at a specific time T_0 of $H\left(T_0\right)$ is represented by two confidence limits which have been constructed so that the probability that the true hazard lies between them is $1 - \alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire hazard function lies within the band. When these are plotted with the hazard curve, these limits must be interpreted on an individual, point by point, basis.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended.

Linear Pointwise Confidence Interval for H(T)

This estimator may be used to create a confidence interval at a specific time point T_0 of $H(T_0)$ using the formula

$$\tilde{H}(T_0) \pm z_{1-\alpha/2}\sigma_{\tilde{H}}(T_0)$$

where *z* is the appropriate value from the standard normal distribution.

Log-Transformed Pointwise Confidence Interval for H(T)

Better confidence limits may be calculated using the logarithmic transformation of the hazard functions. These limits are

$$\tilde{H}(T_0) / \phi, \tilde{H}(T_0)\phi$$

where

$$\phi = \exp\left\{\frac{z_{1-\alpha/2}\sigma_{\tilde{H}}(T_0)}{\tilde{H}(T_0)}\right\}$$

ArcSine-Square Root Pointwise Confidence Interval for H(T)

Another set of confidence limits using an improving transformation is given by the formula

$$-2\ln\left\{\sin\left[\min\left(\frac{\pi}{2}, \arcsin\left[\exp\left\{-\frac{\tilde{H}(T_0)}{2}\right\}\right] + \frac{z_{1-\alpha/2}\sigma_{\tilde{H}}(T_0)}{2\sqrt{\exp\left\{\tilde{H}(T_0)\right\}-1}}\right)\right]\right\}$$

$$\le H(T_0) \le$$

$$-2\ln\left\{\sin\left[\max\left(0, \arcsin\left[\exp\left\{-\frac{\tilde{H}(T_0)}{2}\right\}\right] - \frac{z_{1-\alpha/2}\sigma_{\tilde{H}}(T_0)}{2\sqrt{\exp\left\{\tilde{H}(T_0)\right\}-1}}\right)\right]\right\}$$

Survival Quantiles

The median survival time is an example of a quantile of the survival distribution. It is the smallest value of *T* such that $\hat{S}(T) = 0.50$. In fact, more general results are available for any quantile *p*. The *p*th quantile is estimated by

$$T_p = \inf\left\{T : \hat{S}(T) \le 1 - p\right\}$$

In words, T_p is smallest time at which $\hat{S}(T)$ is less than or equal to 1 - *p*.

A $100(1-\alpha)\%$ confidence interval for T_p can be generated using each of the three estimation methods. These are given next.

Linear Pointwise Confidence Interval for Tp

This confidence interval is given by the set of all times such that

$$-z_{1-\alpha/2} \le \frac{\hat{S}(T)-(1-p)}{\sqrt{\hat{V}\left[\hat{S}(T)\right]}} \le z_{1-\alpha/2}$$

where *z* is the appropriate value from the standard normal distribution.

Log-Transformed Pointwise Confidence Interval for Tp

This confidence interval is given by the set of all times such that

$$-z_{1-\alpha/2} \le \frac{\left[\ln\left\{-\ln\left[\hat{S}(T)\right]\right\}-\ln\left\{-\ln[1-p]\right\}\right]\left[\hat{S}(T)\ln\left[\hat{S}(T)\right]\right]}{\sqrt{\hat{V}\left[\hat{S}(T)\right]}} \le z_{1-\alpha/2}$$

where *z* is the appropriate value from the standard normal distribution.

ArcSine-Square Root Pointwise Confidence Interval for Tp

This confidence interval is given by the set of all times such that

$$-z_{1-\alpha/2} \le \frac{2\left\{\arcsin\left[\sqrt{\hat{S}(T)}\right]-\arcsin\left[\sqrt{1-p}\right]\right\}\sqrt{\hat{S}(T)\left[1-\hat{S}(T)\right]}}{\sqrt{\hat{V}\left[\hat{S}(T)\right]}} \le z_{1-\alpha/2}$$

where *z* is the appropriate value from the standard normal distribution.

Hazard Rate Estimation

The characteristics of the failure process are best understood by studying the hazard rate, $h(T)$, which is the derivative (slope) of the cumulative hazard function $H(T)$. The hazard rate is estimated using kernel smoothing of the Nelson-Aalen estimator as given in Klein and Moeschberger (1997). The formulas for the estimated hazard rate and its variance are given by

$$\hat{h}(T) = \frac{1}{b}\sum_{A \le T_i \le T} K\left(\frac{T-T_i}{b}\right)\Delta\tilde{H}(T_i)$$

$$\sigma^2\left[\hat{h}(T)\right] = \frac{1}{b^2}\sum_{A \le T_i \le T} K\left(\frac{T-T_i}{b}\right)^2 \Delta\hat{V}\left[\tilde{H}(T_i)\right]$$

where b is the bandwidth about T and

$$\Delta\tilde{H}(T_k) = \tilde{H}(T_k) - \tilde{H}(T_{k-1})$$

$$\Delta\hat{V}\left[\tilde{H}(T_k)\right] = \hat{V}\left[\tilde{H}(T_k)\right] - \hat{V}\left[\tilde{H}(T_{k-1})\right]$$

Three choices are available for the kernel function *K*(x) in the above formulation. These are defined differently for various values of *T*. Note that the T_i's are for failed items only and that T_{Max} is the maximum failure time. For the *uniform kernel* the formulas for the various values of *T* are

$$K(x) = \frac{1}{2} \quad for \quad T - b \le T \le T + b$$

$$K_L(x) = \frac{4(1+q^3)}{(1+q)^4} + \frac{6(1-q)}{(1+q)^3}x \quad for \quad T < b$$

$$K_R(x) = \frac{4(1+r^3)}{(1+r)^4} - \frac{6(1-r)}{(1+r)^3}x \quad for \quad T_{Max} - b < T < T_{Max}$$

where

$$q = \frac{T}{b}$$

and

$$r = \frac{T_{Max} - T}{b}$$

For the *Epanechnikov kernel* the formulas for the various values of *T* are

$$K(x) = \frac{3}{4}(1 - x^2) \quad for \quad T - b \le T \le T + b$$

$$K_L(x) = K(x)(A + Bx) \quad for \quad T < b$$

$$K_R(x) = K(-x)(A - Bx) \quad for \quad T_{Max} - b < T < T_{Max}$$

where

$$A = \frac{64(2 - 4q + 6q^2 - 3q^3)}{(1+q)^4(19 - 18q + 3q^2)}$$

$$B = \frac{240(1-q)^2}{(1+q)^4(19 - 18q + 3q^2)}$$

$$q = \frac{T}{b}$$

$$r = \frac{T_{Max} - T}{b}$$

For the *biweight kernel* the formulas for the various values of *T* are

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \quad for \quad T - b \le T \le T + b$$

$$K_L(x) = K(x)(A + Bx) \quad for \quad T < b$$

$$K_R(x) = K(-x)(A - Bx) \quad for \quad T_{Max} - b < T < T_{Max}$$

where

$$A = \frac{64\left(8 - 24q + 48q^2 - 45q^3 + 15q^4\right)}{(1+q)^5\left(81 - 168q + 126q^2 - 40q^3 + 5q^4\right)}$$

$$B = \frac{1120(1-q)^3}{(1+q)^5\left(81 - 168q + 126q^2 - 40q^3 + 5q^4\right)}$$

$$q = \frac{T}{b}$$

$$r = \frac{T_{Max} - T}{b}$$

Confidence intervals for $h(T)$ are given by

$$\hat{h}(T)\exp\left[\pm \frac{z_{1-\alpha/2}\sigma\left[\hat{h}(T)\right]}{\hat{h}(T)}\right]$$

Care must be taken when using this kernel-smoothed estimator since it is actually estimating a smoothed version of the hazard rate, not the hazard rate itself. Thus, it may be biased. Also, it is greatly influenced by the choice of the bandwidth *b*. We have found that you must experiment with *b* to find an appropriate value for each dataset.

Hazard Ratio

Often, it will be useful to compare the hazard rates of two groups. This is most often accomplished by creating the *hazard ratio* (*HR*). The hazard ratio is discussed in depth in Parmar and Machin (1995) and we refer you to this reference for details which we summarize here. The Cox-Mantel estimate of *HR* for two groups A and B is given by

$$HR_{CM} = \frac{H_A}{H_B} = \frac{O_A / E_A}{O_B / E_B}$$

where O_i is the observed number of events (deaths) in group i, E_i is the expected number of events (deaths) in group *i*, and H_i is the overall hazard rate for the ith group. The calculation of the E_i is explained in Parmar and Machin (1995).

A confidence interval for *HR* is found by first transforming to the log scale which is better approximated by the normal distribution, calculating the limits, and then transforming back to the original scale. The calculation is made using

$$\ln\left(HR_{CM}\right) \pm z_{1-\alpha/2}\left(SE_{\ln HR_{CM}}\right)$$

where

$$SE_{\ln HR_{CM}} = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}}$$

which results in the limits

$$\exp\left[\ln\left(HR_{CM}\right) - z_{1-\alpha/2}\left(SE_{\ln HR_{CM}}\right)\right]$$

and

$$\exp\left[\ln\left(HR_{CM}\right) + z_{1-\alpha/2}\left(SE_{\ln HR_{CM}}\right)\right]$$

An alternative estimate of *HR* that is sometimes used is the Mantel-Haenszel estimator which is calculated using

$$HR_{MH} = \exp\left(\frac{O_A - E_A}{V}\right)$$

where *V* is the hypergeometric variance. For further details, see Parmar and Machin (1995). A confidence interval for *HR* is found by first transforming to the log scale which is better approximated by the normal distribution, calculating the limits, and then transforming back to the original scale. The calculation is made using

$$\ln\left(HR_{MH}\right) \pm z_{1-\alpha/2}\left(SE_{\ln HR_{MH}}\right)$$

where

$$SE_{\ln HR_{MH}} = \sqrt{\frac{1}{V}}$$

which results in the limits

$$\exp\left[\ln\left(HR_{MH}\right) - z_{1-\alpha/2}\left(SE_{\ln HR_{MH}}\right)\right]$$

and

$$\exp\left[\ln\left(HR_{MH}\right) + z_{1-\alpha/2}\left(SE_{\ln HR_{MH}}\right)\right]$$

Hypothesis Tests

This section presents methods for testing that the survival curves, and thus the hazard rates, of two or more populations are equal. The specific hypothesis set that is being tested is

$$H_0: h_1(T) = h_2(T) = \cdots = h_K(T), \quad \text{for all } T \le \tau.$$
$$H_A: h_i(T) \ne h_j(T) \text{ for at least one value of } i, j, \text{ and } T \le \tau.$$

Here τ is taken to be the largest observed time in the study.

In words, the null hypothesis is that the hazard rates of all populations are equal at all times less than the maximum observed time and the alternative hypothesis is that at least two of the hazard rates are different at some time less than the observed maximum time.

In the remainder of this section, we will present a general formulation that includes many of the most popular tests. We use the same notation as before, except that now we add an additional subscript, *k*, that represents one of the *K* populations. The test is formed by making a comparison of the actual versus the expected hazard rates. The various hazard rates may be weighted differently. These different weights result in different tests with different properties.

The test is based on the *K*-1 statistics

$$Z_k(\tau) = \sum_{A \le T_i \le T} W(T_i) r_{ik} \left(\frac{d_{ik}}{r_{ik}} - \frac{d_i}{r_i} \right), k = 1, 2, \cdots, K-1$$

where

$$d_i = \sum_{k=1}^{K} d_{ik}$$

$$r_i = \sum_{k=1}^{K} r_{ik}$$

The *Z*'s have a covariance matrix Σ with elements

$$\sigma_{kg} = \sum_{A \le T_i \le T} W(T_i)^2 \left(\frac{r_{ik}}{r_i} \right) \left(\delta_{kg} - \frac{r_{ig}}{r_i} \right) \left(\frac{r_i - d_i}{r_i - 1} \right) d_i$$

where

$$\delta_{kg} = \begin{cases} 1 \text{ if } k = g \\ 0 \text{ if } k \ne g \end{cases}$$

If we let *Z* represent the vector of *K*-1 statistics and Σ represent the covariance matrix, the test statistic is given by

$$Q = Z' \Sigma^{-1} Z$$

In large samples, *Q* is approximately distributed as a chi-squared random variable with *K*-1 degrees of freedom. Details of the above formulas can be found in Klein and Moeschberger (1997), pages 191-202 and Andersen, Borgan, Gill, and Keiding (1992), pages 345-356.

Ten different choices for the weight function, *W*(*T*), result in the ten different tests that are available in ***NCSS***. The most commonly used test is the logrank test which has equal weighting.

The other nine tests shift the heaviest weighting to the beginning or end of the trial. This may be appropriate in some studies, but the use of one of these other weighting schemes should be designated before the data have been seen. Also, even though ten tests are displayed, you should only use one of them. Because of the different weighting patterns, they will often give quite different results. It is bad science to look at all the tests and pick the one that matches your own conclusions. That is why you must designate the test you will use before you have seen the data.

The following table describes each of these tests.

Test	Weight	Comments
Logrank	1	This is the most commonly used test and the one we recommend. Equal weights across all times. This test has optimum power when the hazard rates of the *K* populations are proportional to each other.
Gehan	r_i	Places very heavy weight on hazards at the beginning of the study.
Tarone-Ware	$\sqrt{r_i}$	Places heavy weight on hazards at the beginning of the study.
Peto-Peto	$\tilde{S}(T_i)$	Places a little more weight on hazards at the beginning of the study.
Modified Peto-Peto	$\tilde{S}(T_i)r_i/(r_i+1)$	Places a little more weight on hazards at the beginning of the study.
Fleming-Harrington (0,0)	$1-\hat{S}(T_{i-1})$	Places heavy weight on hazards at the end of the study.
Fleming-Harrington (1,0)	$\hat{S}(T_{i-1})$	Places almost equal weight at all times.
Fleming-Harrington (1,1)	$\hat{S}(T_{i-1})\left(1-\hat{S}(T_{i-1})\right)$	Places heavy weight on hazards at the end of the study.
Fleming-Harrington (0.5,0.5)	$\sqrt{\hat{S}(T_{i-1})\left(1-\hat{S}(T_{i-1})\right)}$	Places a little more weight on hazards at the end of the study.
Fleming-Harrington (0.5,2)	$\left(1-\hat{S}(T_{i-1})\right)^2\sqrt{\hat{S}(T_{i-1})}$	Places very heavy weight on hazards at the end of the study.

This table uses the following definitions.

$$\hat{S}(T)=\prod_{T_i\le T}\left(1-\frac{d_i}{r_i}\right)$$

$$\tilde{S}(T)=\prod_{T_i\le T}\left(1-\frac{d_i}{r_i+1}\right)$$

Logrank Tests

The logrank test is perhaps the most popular test for testing equality of hazard functions. This test uses $W(T) = 1$, that is, equal weighting. This test has optimum power when the hazard rates of the K populations are proportional to each other.

Note that this version of the logrank test is different from the version used in ***NCSS***'s Logrank procedure. That procedure uses the permutation covariance matrix of Lee and Desu which is only valid for equal censoring. The covariance matrix used here is valid for any random censoring pattern, so it is much less restrictive.

Cox-Mantel and Mantel-Haenszel Logrank Tests

When there are only two groups, two versions of the logrank test are commonly used. These tests test the hypothesis that the hazard ratio (HR) is one; that is, that the two hazard rates being compared are zero. Note that these tests are equivalent except in small samples.

Cox-Mantel Logrank Test

Using the notation given above in the section on the hazard ratios, the *Cox-Mantel logrank test* statistic is computed using

$$\chi^2_{CM} = \frac{\left(O_A - E_A\right)^2}{E_A} + \frac{\left(O_B - E_B\right)^2}{E_B}$$

This test statistic is approximately distributed as a chi-square random variable with one degree of freedom.

Mantel-Haenszel Logrank Test

The *Mantel-Haenszel logrank test* statistic is computed using

$$\chi^2_{CM} = \frac{\left(O_A - E_A\right)^2}{V}$$

This test statistic is also approximately distributed as a chi-square random variable with one degree of freedom.

Randomization Probability Levels

Because of assumptions that must be made when using this procedure, ***NCSS*** also includes a randomization test as outlined by Edgington (1987). Randomization tests are becoming more and more popular as the speed of computers allows them to be computed in seconds rather than hours.

A randomization test is conducted by forming a Monte Carlo sampling of all possible permutations of the sample data, calculating the test statistic for each sampled permutation, and counting the number of permutations that result in a chi-square value greater than or equal to the actual chi-square value. Dividing this count by the number of permutations sampled gives the significance level of the test. Edgington suggests that at least 1,000 permutations by selected.

Data Structure

Survival data sets require up to three components for the survival time: the ending survival time, an optional beginning survival time during which the subject was not observed, and an indicator of whether an observation was censored or failed.

Based on these three components, various types of data may be analyzed. Right censored data are specified using only the ending time variable and the censor variable. Left truncated and Interval data are entered using all three variables.

Sample Dataset

Most survival data sets require at least two variables: the failure time variable and a censor variable that indicates whether time is a failure or was censored. Optional variables include a count variable which gives the number of items occurring at that time and a group variable that identifies which group this observation belongs to. If the censor variable is omitted, all time values represent failed items. If the count variable is omitted, all counts are assumed to be one.

The table below shows a dataset reporting on a two-group time-to-tumor study. In this data set, time-to-tumor (in days) is given for twelve mice. The twelve mice were randomly divided into two groups. The first group served as a control group, while the second group received a dose of a certain chemical. These data are contained in the SURVIVAL database.

SURVIVAL dataset (subset)

Tumor6	Censor6	Trtmnt6
8	1	1
8	1	1
10	1	1
12	1	1
12	1	1
13	1	1
9	1	2
12	1	2
15	1	2
20	1	2
30	0	2
30	0	2

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Variables

Time Variable

This variable contains the length of time that an individual was observed. This may represent a failure time or a censor time. Whether the subject actually died is specified by the Censor Variable. Since the values are elapsed times, they must be positive. Zeroes and negative values are treated as missing values.

During the maximum likelihood calculations, a risk set is defined for each individual. The risk set is defined to be those subjects who were being observed at this subject's failure and who lived as long or longer. It may take several rows of data to specify a subject's history.

This variable and the Entry Time Variable define a period during which the individual was at risk of failing. If the Entry Time Variable is not specified, its value is assumed to be zero.

Several types of data may be entered. These will be explained next.

- **Failure**

  This type of data occurs when a subject is followed from their entrance into the study until their death. The failure time is entered in this variable and the Censor Variable is set to the failed code, which is often a one.

  The Entry Time Variable is not necessary. If an Entry Time Variable is used, its value should be zero for this type of observation.

- **Interval Failure**

  This type of data occurs when a subject is known to have died during a certain interval. The subject may, or may not, have been observed during other intervals. If they were, they are treated as Interval Censored data. An individual may require several rows on the database to record their complete follow-up history.

  For example, suppose the condition of the subjects is only available at the end of each month. If a subject fails during the fifth month, two rows of data would be required. One row, representing the failure, would have a Time of 5.0 and an Entry Time of 4.0. The Censor variable would contain the failure code. A second row, representing the prior periods, would have a Time of 4.0 and an Entry Time of 0.0. The Censor variable would contain the censor code.

- **Right Censored**

  This type of data occurs when a subject has not failed up to the specified time. For example, suppose that a subject enters the study and does not die until after the study ends 12 months later. The subject's time (365 days) is entered here. The Censor variable contains the censor code.

- **Interval Censored**

  This type of data occurs when a subject is known not to have died during a certain interval. The subject may, or may not, have been observed during other intervals. An individual may require several rows on the database to record their complete follow-up history.

  For example, suppose the condition of the subjects is only available at the end of each month. If a subject fails during the fifth month, two rows of data would be required. One row, representing the failure, would have a Time of 5.0 and an Entry Time of 4.0. The Censor variable would contain the failure code. A second row, representing the prior periods, would have a Time of 4.0 and an Entry Time of 0.0. The Censor variable would contain the censor code.

Entry Time Variable

This optional variable contains the elapsed time before an individual entered the study. Usually, this value is zero. However, in cases such as *left truncation* and *interval censoring*, this value defines a time period before which the individual was not observed.

Negative entry times are treated as missing values. It is possible for the entry time to be zero.

Min Entry Time

When you have left truncation, this value gives the minimum entry time after which events are considered. When used, all survival and hazard rates are conditional on the subject reaching this age.When there is no left truncation (Entry Time Variable), this value is set to zero.

This value is necessary because with left truncation, it is possible for the Kaplan-Meier estimate to reach (and then stay at) zero to soon. Conditioning the probability statements so that this age must be reached in order for an individual to be in a risk set removes this zeroing out problem.

Censor Variable

Censor Variable

The values in this variable indicate whether the value of the Time Variable represents a censored time or a failure time. These values may be text or numeric. The interpretation of these codes is specified by the Failed and Censored options to the right of this option.

Only two values are used, the Failure code and the Censor code. The Unknown Type option specifies what is to be done with values that do not match either the Failure code or the Censor code.

Rows with missing values (blanks) in this variable are omitted.

Failed

This value identifies those values of the Censor Variable that indicate that the Time Variable gives a failure time. The value may be a number or a letter.

We suggest the letter ‘F’ or the number ‘1’ when you are in doubt as to what to use.

A failed observation is one in which the time until the event of interest was measured exactly; for example, the subject died of the disease being studied. The exact failure time is known.

(Left Censoring)

When the exact failure time is not known, but instead only an upper bound on the failure time is known, the time value is said to have been *left censored*. In this case, the time value is treated as if it were the true failure time, not just an upper bound. So left censored observations should be coded as failed observations.

Censored

This value identifies those values of the Censor Variable that indicate that the individual recorded on this row was censored. That is, the actual failure time occurs sometime after the value of the Time Variable.

We suggest the letter ‘C’ or the number ‘0’ when you are in doubt as to what to use.

A censored observation is one in which the time until the event of interest is not known because the individual withdrew from the study, the study ended before the individual failed, or for some similar reason.

Note that it does not matter whether the censoring was Right or Interval. All you need to indicate here is that they were censored.

Unknown Censor

This option specifies what the program is to assume about rows whose censor value is not equal to either the Failed code or the Censored code. Note that observations with missing censor values are always treated as missing.

- **Censored**
  Observations with unknown censor values are assumed to have been censored.
- **Failed**
  Observations with unknown censor values are assumed to have failed.
- **Missing**
  Observations with unknown censor values are assumed to be missing and they are removed from the analysis.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable and the log rank tests are performed. If it is not specified, the log rank tests cannot be generated.

Resampling

Run Randomization tests

Check this option to run randomization tests. Note that these tests are computer-intensive and may require a great deal of time to run.

Monte Carlo Samples

Specify the number of Monte Carlo samples used when conducting randomization tests. You also need to check the 'Run randomization tests' box to run these tests.

Somewhere between 1,000 and 100,000 Monte Carlo samples are usually necessary. We suggest the use of 10,000.

Frequency Variable

Frequency Variable

This variable gives the count, or frequency, of the time displayed on that row. When omitted, each row receives a frequency of one. Frequency values should be positive integers. This is usually used to indicate the number of right censored values at the end of a study or the number of failures occurring within an interval. It may also be used to indicate ties for failure data.

Options – KM Survival and Cumulative Hazard

Confidence Limits

This option specifies the method used to estimate the confidence limits of the Kaplan-Meier Survival and the Cumulative Hazard. The options are:

- **Linear**

  This is the classical method which uses Greenwood's estimate of the variance.

- **Log Transform**

  This method uses the logarithmic transformation of Greenwood's variance estimate. It produces better limits than the Linear method and has better small sample properties.

- **ArcSine**

  This method uses the arcsine square-root transformation of Greenwood's variance estimate to produce better limits.

Variance

The option specifies which estimator of the variance of the Nelson-Aalen cumulative hazard estimate is to be used. Three estimators have been proposed. When there are no event-time ties, all three give about the same results.

We recommend that you use the Simple estimator unless ties occur naturally in the theoretical event times.

- **Simple**

  This estimator should be used when event-time ties are caused by rounding and lack of measurement precision. This estimate gives the largest value and hence the widest, most conservation, confidence intervals.

- **Plug In**

  This estimator should be used when event-time ties are caused by rounding and lack of measurement precision.

- **Binomial**

  This estimator should be used when ties occur in the theoretical distribution of event times.

Options – Hazard Rate

The following options control the calculation of the hazard rate and cumulative hazard function.

Bandwidth Method

This option designates the method used to specify the smoothing bandwidth used to calculate the hazard rate. Specify an amount or a percentage of the time range. The default is to specify a percent of the time range.

Bandwidth Amount

This option specifies the bandwidth size used to calculate the hazard rate. If the Bandwidth Method was set to Amount, this is a value in time units (such as 10 hours). If Percentage of Time Range was selected, this is a percentage of the overall range of the data.

Smoothing Kernel

This option specifies the kernel function used in the smoothing to estimate the hazard rate. You can select uniform, Epanechnikov, or biweight smoothing. The actual formulas for these functions were provided above.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Section - Hazard Ratio Tests

These options indicate whether to display the corresponding report.

Report Options

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. Single precision is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, only value labels, or both for values of the group variable. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Male, 2=Female, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Survival and Hazard Rate Calculation Values

Percentiles

This option specifies a list of percentiles (range 1 to 99) at which the reliability (survivorship) is reported. The values should be separated by commas.

You can specify sequences with a colon, putting the increment inside parentheses after the maximum in the sequence. For example: 5:25(5) means 5,10,15,20,25 and 1:5(2),10:20(2) means 1,3,5,10,12,14,16,18,20.

Times

This option specifies a list of times at which the percent surviving and cumulative hazard values are reported. Individual values are separated by commas. You can specify a sequence by specifying the minimum and maximum separate by a colon and putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers. Use '(10)' alone to specify ten values between zero and the maximum value found in the data.

Report Options – Decimal Places

Time

This option specifies the number of decimal places shown on reported time values.

Probability

This option specifies the number of decimal places shown on reported probability and hazard values.

Chi-Square

This option specifies the number of decimal places shown on reported chi-square values.

Ratio

This option specifies the number of decimal places shown on reported ratio values.

Plots Tab

The following options control the plots that are displayed.

Select Plots

These options specify which plots type of plots are displayed.

Survival/Reliability Plot – Hazard Rate Plot

Specify whether to display each of these plots.

Select Plots – Plots Displayed

Individual-Group Plots

When checked, this option specifies that a separate chart of each designated type is displayed.

Combined Plots

When checked, this option specifies that a chart combining all groups is to be displayed.

Plot Options – Plot Arrangement

These options control the size and arrangement of the plots.

Two Plots Per Line

When a man charts are specified, checking this option will cause the size of the charts to be reduced so that they can be displayed two per line. This will reduce the overall size of the output.

Plot Options – Plot Contents

These options control objects that are displayed on all plots.

Function Line

Indicate whether to display the estimated survival (Kaplan-Meier) or hazard function on the plots.

C.L. Lines

Indicate whether to display the confidence limits of the estimated function on the plots.

Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the name of the group variable.

Horizontal (Time) Axis

These options control the horizontal axis of the plots.

Label

This is the text of the horizontal label. The characters *{X}* are replaced the name of the time variable. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the horizontal (X) axes. If left blank, these values are calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Survival Plot Tab

These options control the attributes of the survival curves. Note that the horizontal axis is specified in the Plots tab.

Vertical Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Survival Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Censor Tickmarks

This option indicates the size of the tickmarks (if any) showing where the censored points fall on the Kaplan-Meier survival curve. The values are at a scale of 1000 equals one inch.

We recommend that you use ‘0’ to indicate no marks or ‘100’ to display the marks.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, and *{Z}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Cum Haz Plot Tab

These options control the attributes of the cumulative hazard function plot.

Vertical Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Cum Hazard Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, and *{Z}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Haz Rt Plot Tab

These options control the attributes of the hazard rate plot.

Vertical Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Hazard Rate Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, and *{Z}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines Tab

These options specify the attributes of the lines used for each group in the hazard curves and survival curves.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Labels Tab

The options on this tab specify the labels that a printed on the reports and plots.

Report and Plot Labels

Failure Time Label - Hazard Rate Label

These options specify the term(s) used as labels for these items on the plots. Since these reports are used for performing survival analysis in medical research and reliability analysis in industry, and since these fields often use different terminology, these options are needed to provide appropriate headings for the reports.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful.

Data Storage Options – Select Items to Store on the Spreadsheet

Survival Group - Hazard Rate UCL

Indicate whether to store these values, beginning at the variable indicated by the *Store First Variable In* option.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Kaplan-Meier Survival Analysis

This section presents an example of how to analyze a typical set of survival data. In this study, thirty subjects were watched to see how long until a certain event happened after the subject received a certain treatment. The study was terminated at 152.7 hours. At this time, the event had not occurred in eighteen of the subjects. The data used are recorded in the WEIBULL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Kaplan-Meier Curves (Logrank Tests) Fitting window.

1 Open the WEIBULL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **WEIBULL.S0**.
- Click **Open**.

2 Open the Kaplan-Meier Curves (Logrank Tests) window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Kaplan-Meier Curves (Logrank Tests)**. The Kaplan-Meier Curves (Logrank Tests) procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Variables tab**.
- Set the **Time Variable** to **Time**.
- Set the **Censor Variable** to **Censor**.
- Set the **Frequency Variable** to **Count**.

4 Specify the reports.

- Select the **Reports tab**.
- Set the **Times** box to **10:150(10)**.

5 Specify the plots.

- Select the **Plots tab**.
- Check **Hazard Function Plot**.
- Check **Hazard Rate Plot**.
- Check **C.L. Lines**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum
Failed	12	12	12.5	152.7
Censored	1	18	152.7	152.7
Total	13	30	12.5	152.7

This report displays a summary of the amount of data that were analyzed. Scan this report to determine if there were any obvious data errors by double checking the counts and the minimum and maximum times.

Survival at Specific Event Times

Event Time (T)	Cumulative Survival S(T)	Standard Error of S(T)	Lower 95% C.L. for S(T)	Upper % C.L. for S(T)	At Risk
10.0	1.0000	0.0000	1.0000	1.0000	30
20.0	0.9667	0.0328	0.9024	1.0000	29
30.0	0.9333	0.0455	0.8441	1.0000	28
40.0	0.9333	0.0455	0.8441	1.0000	28
50.0	0.9333	0.0455	0.8441	1.0000	28
60.0	0.9000	0.0548	0.7926	1.0000	27
70.0	0.8333	0.0680	0.7000	0.9667	25
80.0	0.8333	0.0680	0.7000	0.9667	25
90.0	0.8333	0.0680	0.7000	0.9667	25
100.0	0.7333	0.0807	0.5751	0.8916	22
110.0	0.7333	0.0807	0.5751	0.8916	22
120.0	0.7000	0.0837	0.5360	0.8640	21
130.0	0.6333	0.0880	0.4609	0.8058	19
140.0	0.6333	0.0880	0.4609	0.8058	19
150.0	0.6333	0.0880	0.4609	0.8058	19

This report displays the Kaplan-Meier product-limit survival probabilities at the specified time points. The formulas used were presented earlier.

Event Time (T)

This is the time point being reported on this line. The time values were specified in the Times box under the Report tab.

Cumulative Survival S(T)

This is the probability that a subject does not have the event until after the event time given on this line. This probability is estimated using the Kaplan-Meier product limit method. The estimate is given by the formula

$$\hat{S}(T) = \begin{cases} 1 & \text{if } T_{\min} > T \\ \prod_{A \le T_i \le T} \left[1 - \frac{d_i}{r_i} \right] & \text{if } T_{\min} \le T \end{cases}$$

Standard Error of S(T)

This is the estimated standard error of the Kaplan-Meier survival probability. The variance of *S(T)* is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(T)\right] = \hat{S}(T)^2 \sum_{A \le T_i \le T} \frac{d_i}{r_i(r_i - d_i)}$$

The standard error is the square root of this variance.

Lower and Upper Confidence Limits for S(T)

The lower and upper confidence limits provide a pointwise confidence interval for the survival probability at each time point. These limits are constructed so that the probability that the true survival probability lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire survival function lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

At Risk

This value is the number of individuals at risk. The number at risk is all those under study who died or were censored at a time later than the current time. As the number of individuals at risk is decreased, the estimates become less reliable.

Quantiles of Survival Time

Proportion Surviving	Proportion Failing	Survival Time	Lower 95% C.L. Survival Time	Upper 95% C.L. Survival Time
0.9500	0.0500	24.4	12.5	69.1
0.9000	0.1000	58.2	24.4	96.6
0.8500	0.1500	69.1	24.4	114.2
0.8000	0.2000	95.5	58.2	125.6
0.7500	0.2500	97.0	68.0	152.7
0.7000	0.3000	114.2	69.1	152.7
0.6500	0.3500	125.6	96.6	152.7
0.6000	0.4000	152.7	97.0	152.7
0.5500	0.4500		114.2	152.7
0.5000	0.5000		123.2	152.7
0.4500	0.5500		152.7	152.7
0.4000	0.6000			152.7
0.3500	0.6500			152.7
0.3000	0.7000			152.7
0.2500	0.7500			152.7
0.2000	0.8000			152.7
0.1500	0.8500			152.7
0.1000	0.9000			152.7
0.0500	0.9500			152.7

This report displays the estimated survival times for various survival proportions. For example, it gives the median survival time if it can be estimated.

Proportion Surviving

This is the proportion surviving that is reported on this line. The proportion values were specified in the Percentiles box under the Report tab.

Proportion Failing

This is the proportion failing. The proportion is equal to one minus the proportion surviving.

Survival Time

This is the time value corresponding to the proportion surviving. The *p*th quantile is estimated by

$$T_p = \inf\left\{T : \hat{S}(T) \le 1 - p\right\}$$

In words, T_p is smallest time at which $\hat{S}(T)$ is less than or equal to 1 - *p*.

For example, this table estimates that 95% of the subjects will survive longer than 24.4 hours.

Lower and Upper Confidence Limits on Survival Time

These values provide a pointwise $100(1-\alpha)\%$ confidence interval for T_p. For example, if *p* is 0.50, this provides a confidence interval for the median survival time.

Three methods are available for calculating these confidence limits. The method is designated under the Variables tab in the Confidence Limits box. The formulas for these confidence limits were given in the Survival Quantiles section.

Note that because of censoring, estimates and confidence limits are not available for all survival proportions.

Cumulative Hazards for Specific Times

Event Time (T)	Cumulative Hazard H(T)	Standard Error of H(T)	Lower 95% C.L. for H(T)	Upper 95% C.L. for H(T)	At Risk
10.0	0.0000	0.0000	0.0000	0.0000	30
20.0	0.0333	0.0333	0.0000	0.0987	29
30.0	0.0678	0.0480	0.0000	0.1618	28
40.0	0.0678	0.0480	0.0000	0.1618	28
50.0	0.0678	0.0480	0.0000	0.1618	28
60.0	0.1035	0.0598	0.0000	0.2207	27
70.0	0.1790	0.0802	0.0219	0.3362	25
80.0	0.1790	0.0802	0.0219	0.3362	25
90.0	0.1790	0.0802	0.0219	0.3362	25
100.0	0.3042	0.1079	0.0926	0.5158	22
110.0	0.3042	0.1079	0.0926	0.5158	22
120.0	0.3496	0.1171	0.1201	0.5792	21
130.0	0.4472	0.1360	0.1808	0.7137	19
140.0	0.4472	0.1360	0.1808	0.7137	19
150.0	0.4472	0.1360	0.1808	0.7137	19

This report displays estimates of the cumulative hazard function at the specified time points. The formulas used were presented earlier.

Event Time (T)

This is the time point being reported on this line. The time values were specified in the Times box under the Report tab.

Cumulative Hazard H(T)

This is the Nelson-Aalen estimator of the cumulative hazard function, $H(T)$. This estimator is give by

$$\tilde{H}(T) = \begin{cases} 0 & if\ T_{\min} > T \\ \sum_{A \le T_i \le T} \frac{d_i}{r_i} & if\ T_{\min} \le T \end{cases}$$

Standard Error of H(T)

This is the estimated standard error of the above cumulative hazard function. The formula used was specified under the Variables tab in the Variance box. These formulas were given above in the section discussing the Nelson-Aalen estimator.

The standard error is the square root of this variance.

Lower and Upper Confidence Limits for H(T)

The lower and upper confidence limits provide a pointwise confidence interval for the cumulative hazard at each time point. These limits are constructed so that the probability that the true cumulative hazard lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire cumulative hazard function lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

At Risk

This value is the number of individuals at risk. The number at risk is all those under study who died or were censored at a time later than the current time. As the number of individuals at risk is decreased, the estimates become less reliable.

Hazard Rates Section

Failure Time	Nonparametric Hazard Rate	Std Error of Hazard Rate	95% Lower Conf. Limit of Hazard Rate	95% Upper Conf. Limit of Hazard Rate
10.0	0.0018	0.0014	0.0004	0.0085
20.0	0.0018	0.0013	0.0005	0.0073
30.0	0.0015	0.0010	0.0004	0.0055
40.0	0.0016	0.0009	0.0005	0.0047
50.0	0.0022	0.0012	0.0008	0.0063
60.0	0.0026	0.0015	0.0008	0.0080
70.0	0.0034	0.0016	0.0014	0.0084
80.0	0.0042	0.0017	0.0019	0.0095
90.0	0.0043	0.0019	0.0018	0.0101
100.0	0.0048	0.0021	0.0020	0.0111
110.0	0.0054	0.0022	0.0024	0.0121
120.0	0.0047	0.0021	0.0019	0.0113
130.0	0.0038	0.0020	0.0014	0.0105
140.0	0.0036	0.0025	0.0009	0.0143
150.0	0.0066	0.0066	0.0009	0.0468

This report displays estimates of the hazard rates at the specified time points. The formulas used were presented earlier.

Failure Time

This is the time point being reported on this line. The time values were specified in the Times box under the Report tab.

Nonparametric Hazard Rate

The characteristics of the failure process are best understood by studying the hazard rate, $h(T)$, which is the derivative (slope) of the cumulative hazard function $H(T)$. The hazard rate is estimated using kernel smoothing of the Nelson-Aalen estimator as given in Klein and Moeschberger (1997). The formulas used were given earlier and are not repeated here.

Care must be taken when using this kernel-smoothed estimator since it is actually estimating a smoothed version of the hazard rate, not the hazard rate itself. Thus, it may be biased. Also, it is greatly influenced by the choice of the bandwidth b. We have found that you must experiment with b to find an appropriate value for each dataset.

The values of the smoothing parameters are specified under the Hazards tab.

Standard Error of Hazard Rate

This is the estimated standard error of the above hazard rate. The formula used was specified under the Variables tab in the Variance box. These formulas were given above in the section discussing the Nelson-Aalen estimator.

The standard error is the square root of this variance.

Lower and Upper Confidence Limits of Hazard Rate

The lower and upper confidence limits provide a pointwise confidence interval for the smoothed hazard rate at each time point. These limits are constructed so that the probability that the true hazard rate lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire hazard rate function lies within the band.

Product-Limit Survival Analysis

Event Time (T)	Cumulative Survival S(T)	Standard Error of S(T)	Lower 95% C.L. for S(T)	Upper 95% C.L. for S(T)	At Risk	Count	Total Events
12.5	0.9667	0.0328	0.9024	1.0000	30	1	1
24.4	0.9333	0.0455	0.8441	1.0000	29	1	2
58.2	0.9000	0.0548	0.7926	1.0000	28	1	3
68.0	0.8667	0.0621	0.7450	0.9883	27	1	4
69.1	0.8333	0.0680	0.7000	0.9667	26	1	5
95.5	0.8000	0.0730	0.6569	0.9431	25	1	6
96.6	0.7667	0.0772	0.6153	0.9180	24	1	7
97.0	0.7333	0.0807	0.5751	0.8916	23	1	8
114.2	0.7000	0.0837	0.5360	0.8640	22	1	9
123.2	0.6667	0.0861	0.4980	0.8354	21	1	10
125.6	0.6333	0.0880	0.4609	0.8058	20	1	11
152.7	0.6000	0.0894	0.4247	0.7753	19	1	12
152.7+					18	18	12

This report displays the Kaplan-Meier product-limit survival distribution along with confidence limits. The formulas used were presented earlier.

Also note that the sample size is given for each time period. As time progresses, participants are removed from the study, reducing the sample size. Hence, the survival results near the end of the study are based on only a few participants and are therefore less precise. This shows up as a widening of the confidence limits.

Event Time (T)

This is the time point being reported on this line. The time values are specific event times that occurred in the data.

Note that censored observations are marked with a plus sign on their time value. The survival functions are not calculated for censored observations.

Cumulative Survival S(T)

This is the probability that a subject does not have the event until after the event time given on this line. This probability is estimated using the Kaplan-Meier product limit method. The estimate is given by the formula

$$\hat{S}(T) = \begin{cases} 1 & \text{if } T_{\min} > T \\ \prod_{A \le T_i \le T} \left[1 - \frac{d_i}{r_i}\right] & \text{if } T_{\min} \le T \end{cases}$$

Standard Error of S(T)

This is the estimated standard error of the Kaplan-Meier survival probability. The variance of *S(T)* is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(T)\right] = \hat{S}(T)^2 \sum_{A \le T_i \le T} \frac{d_i}{r_i(r_i - d_i)}$$

The standard error is the square root of this variance.

Lower and Upper Confidence Limits for S(T)

The lower and upper confidence limits provide a pointwise confidence interval for the survival probability at each time point. These limits are constructed so that the probability that the true survival probability lies between them is $1 - \alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire survival function lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

At Risk

This value is the number of individuals at risk. The number at risk is all those under study who died or were censored at a time later than the current time. As the number of individuals at risk is decreased, the estimates become less reliable.

Count

This is the number of individuals having the event (failing) at this time point.

Total Events

This is the cumulative number of individuals having the event (failing) up to and including this time value.

Nelson-Aalen Cumulative Hazard Section

Event Time (T)	Cumulative Hazard H(T)	Standard Error of H(T)	Lower 95% C.L. for H(T)	Upper 95% C.L. for H(T)	At Risk	Count	Total Events
12.5	0.0333	0.0333	0.0000	0.0987	30	1	1
24.4	0.0678	0.0480	0.0000	0.1618	29	1	2
58.2	0.1035	0.0598	0.0000	0.2207	28	1	3
68.0	0.1406	0.0703	0.0027	0.2784	27	1	4
69.1	0.1790	0.0802	0.0219	0.3362	26	1	5
95.5	0.2190	0.0896	0.0434	0.3946	25	1	6
96.6	0.2607	0.0988	0.0670	0.4544	24	1	7
97.0	0.3042	0.1079	0.0926	0.5158	23	1	8
114.2	0.3496	0.1171	0.1201	0.5792	22	1	9
123.2	0.3972	0.1264	0.1494	0.6451	21	1	10
125.6	0.4472	0.1360	0.1808	0.7137	20	1	11
152.7	0.4999	0.1458	0.2141	0.7856	19	1	12
152.7+					18	18	12

This report displays estimates of the cumulative hazard function at the time points encountered in the dataset. The formulas used were presented earlier.

Event Time (T)

This is the time point being reported on this line. The time values are specific event times that occurred in the data.

Note that censored observations are marked with a plus sign on their time value. The survival functions are not calculated for censored observations.

Cumulative Hazard H(T)

This is the Nelson-Aalen estimator of the cumulative hazard function, $H(T)$.

Standard Error of H(T)

This is the estimated standard error of the above cumulative hazard function. The formula used was specified under the Variables tab in the Variance box. These formulas were given above in the section discussing the Nelson-Aalen estimator.

The standard error is the square root of this variance.

Lower and Upper Confidence Limits for H(T)

The lower and upper confidence limits provide a pointwise confidence interval for the cumulative hazard at each time point. These limits are constructed so that the probability that the true cumulative hazard lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire cumulative hazard function lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

At Risk

This value is the number of individuals at risk. The number at risk is all those under study who died or were censored at a time later than the current time. As the number of individuals at risk is decreased, the estimates become less reliable.

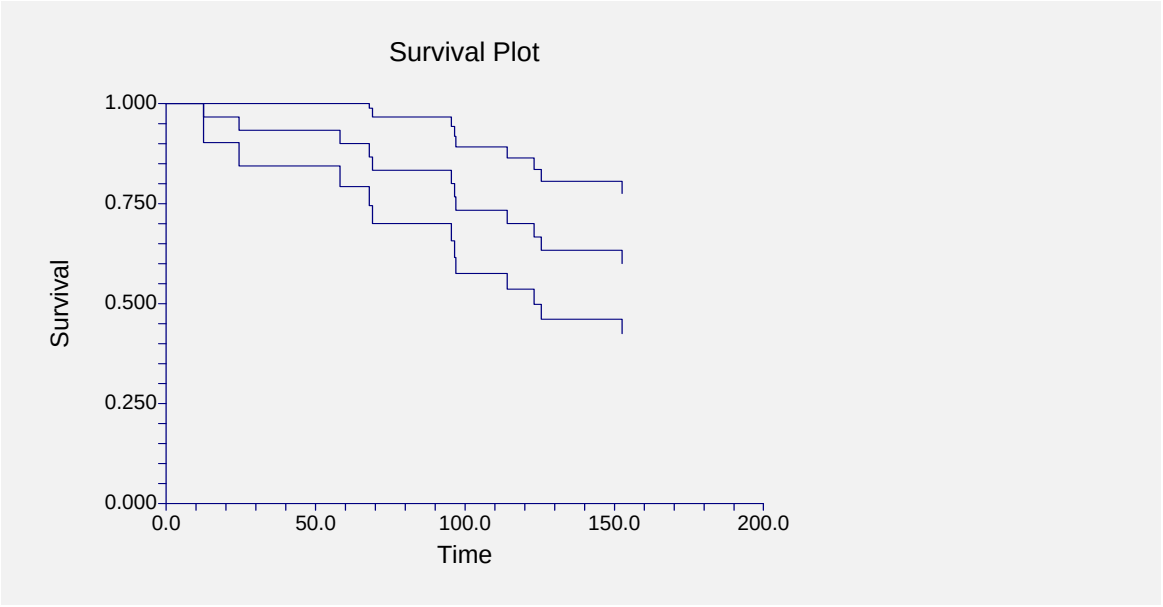
Count

This is the number of individuals having the event (failing) at this time point.

Total Events

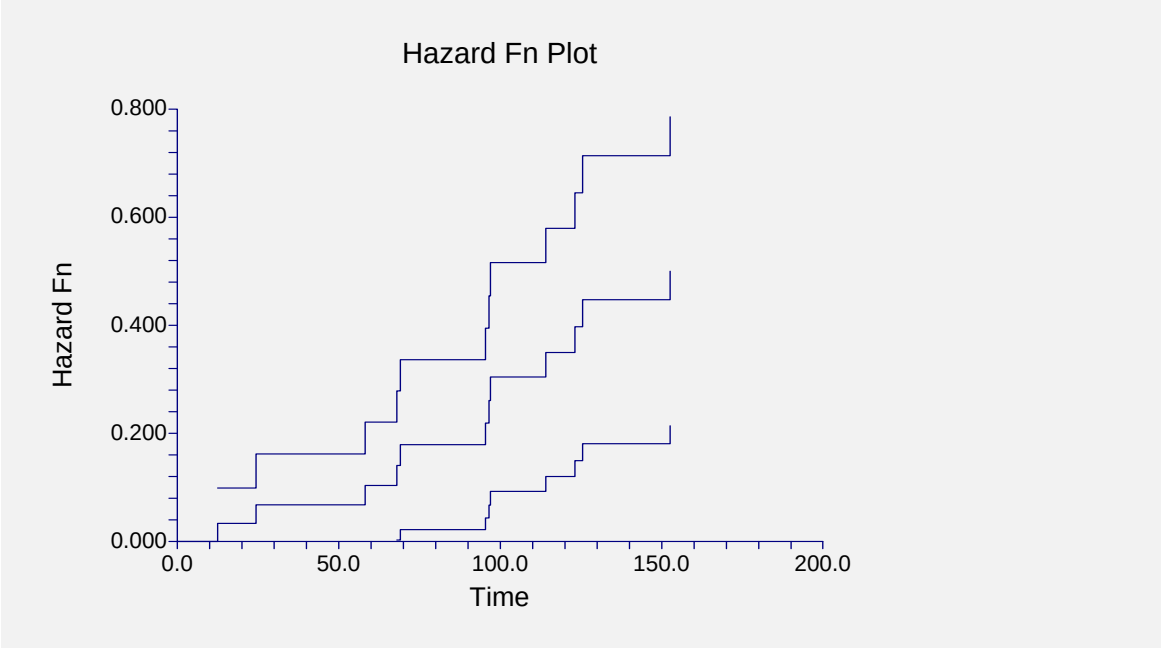
This is the cumulative number of individuals having the event (failing) up to and including this time value.

Survival Plot



This plot shows the product-limit survivorship function as well as the pointwise confidence intervals. If there are several groups, a separate line is drawn for each group.

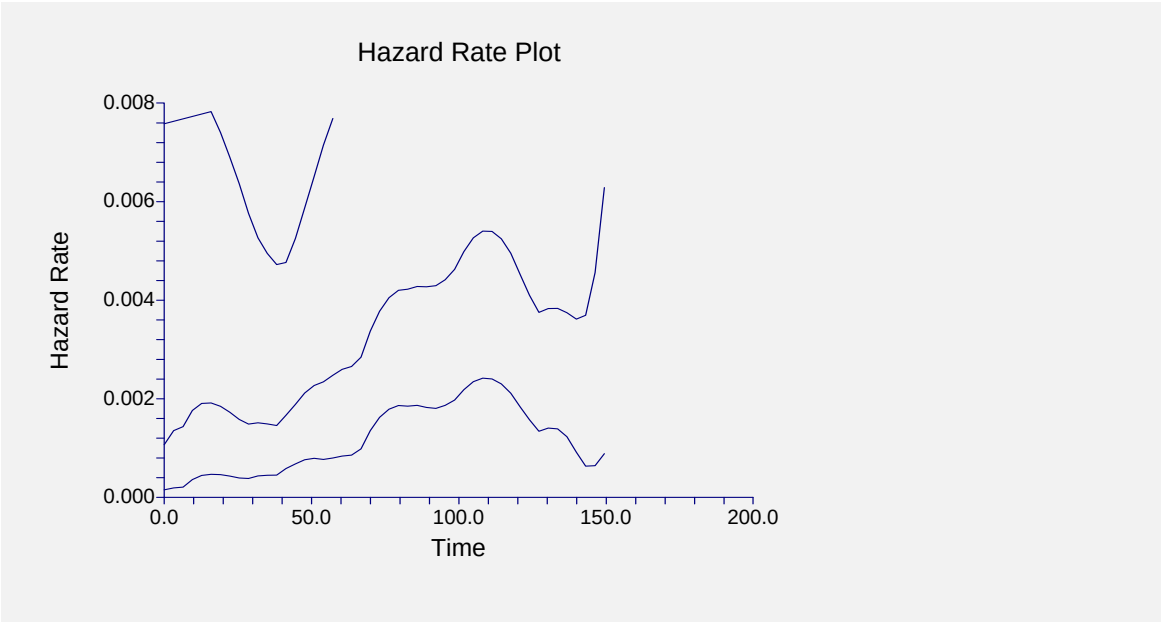
Hazard Function Plot



This plot shows the Nelson-Aalen cumulative hazard function for the data analyzed. Confidence limits are also given.

If you have several groups, then a separate line is drawn for each group.

Hazard Rate Plot



This plot shows the hazard rate with associated confidence limits.

Example 2 – Logrank Tests

This section presents an example of how to use a logrank test to compare the hazard rates of two or more groups.

The data used are recorded in the variables Tumor6, Censor6, and Trtmnt6 of the SURVIVAL database.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Kaplan-Meier Curves (Logrank Tests) Fitting window.

1 Open the SURVIVAL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SURVIVAL.S0**.
- Click **Open**.

2 Open the Kaplan-Meier Curves (Logrank Tests) window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Kaplan-Meier Curves (Logrank Tests)**. The Kaplan-Meier Curves (Logrank Tests) procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Variables tab**.
- Set the **Time Variable** to **TUMOR6**.
- Set the **Censor Variable** to **CENSOR6**.
- Set the **Group Variable** to **TRTMNT6**.
- Check the **Run Randomization Tests** box.
- Set the **Monte Carlo Samples** to **1000**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Logrank Tests Section

Test Name	Chi-Square	DF	Prob Level*	Randomization Test Prob Level*	Weighting of Hazard Comparisons Across Time
Logrank	4.996	1	0.0254	0.0470	Equal
Gehan-Wilcoxon	3.956	1	0.0467	0.0520	High++ to Low++
Tarone-Ware	4.437	1	0.0352	0.0510	High to Low
Peto-Peto	3.729	1	0.0535	0.0670	High+ to Low+
Mod. Peto-Peto	3.618	1	0.0572	0.0740	High+ to Low+
F-H (1, 0)	3.956	1	0.0467	0.0520	Almost Equal
F-H (.5, .5)	3.507	1	0.0611	0.0920	Low+ to High+
F-H (1, 1)	4.024	1	0.0449	0.0760	Low to High
F-H (0, 1)	5.212	1	0.0224	0.0650	Low to High
F-H (.5, 2)	5.942	1	0.0148	0.0420	Low++ to High++

*These probability levels are only valid when a single test was selected before this analysis is seen.
You cannot select a test to report after viewing this table without adding bias to the results.
Unless you have good reason to do otherwise, you should use the equal weighting (logrank) test.

This report gives the results of the ten logrank type tests that are provided by this procedure. We strongly suggest that you select the test that will be used before viewing this report. Unless you have a good reason for doing so, we recommend that you use the first (Logrank) test.

Chi-Square

This is the chi-square value of the test. Each of these tests is approximately distributed as a chi-square in large samples.

DF

This is the degrees of freedom of the chi-square distribution. It is one less than the number of groups.

Prob Level

This is the significance level of the test. If this value is less than your chosen significance level (often 0.05), the test is significant and the hazard rates of the groups are not identical at all time values.

Randomization Test Prob Level

This is the significance level of the corresponding randomization test. This significance level is exact if the assumption that any censoring is independent of which group the subject was in.

In this example, several of the tests that were just significant at the 0.05 level are not significant using corresponding the randomization test. In cases like this, the randomization test should be considered more accurate than the chi-square test.

Weighting of Hazard Comparisons Across Time

The type of weighting pattern that is used by this test is given here.

Logrank Test Detail Section

Logrank Test Detail Section

Group	Z-Value	Standard Error	Standardized Z-Value
1	2.831	1.266	2.235
2	-2.831	1.266	-2.235

Probability Level was 0.0254

Gehan-Wilcoxon Test Detail Section

Group	Z-Value	Standard Error	Standardized Z-Value
1	24.000	12.066	1.989
2	-24.000	12.066	-1.989

Probability Level was 0.0467

Tarone-Ware Test Detail Section

Group	Z-Value	Standard Error	Standardized Z-Value
1	8.130	3.859	2.106
2	-8.130	3.859	-2.106

Probability Level was 0.0352

.
.
.
report continues for all ten tests.

This report gives the details of each of the ten logrank tests that are provided by this procedure. We strongly suggest that you select the test that will be used before viewing this report. Unless you have a good reason for doing so, we recommend that you use the first (Logrank) test.

Group

This is the group reported about on this line.

Z-Value

This is a weighted average of the difference between the observed hazard rates of this group and the expected hazard rates under the null hypothesis of hazard rate equality. The expected hazard rates are found by computing new hazard rates based on all that data as if they all came from a single group.

By considering the magnitudes of these values, you can determine which group (or groups) are different from the rest.

Standard Error

This is the standard error of the above z-value. It is used to standardize the z-values.

Standardized Z-Value

The standardized z-value is created by dividing the z-value by its standard error. This provides an index number that will usually very between -3 and 3. Larger values represent groups that quite different from the typical group, at least at some time values.

Hazard Ratio Detail Section

Groups	Sample Size (nA/nB)	Observed Events (OA/OB)	Expected Events (EA/EB)	Hazard Rates (HRA/HRB)	Cox-Mantel Hazard Ratio (HR)	Hyper. Var. (V)
1/2	6/6	6/4	3.17/6.83	1.89/0.59	3.23	1.60
2/1	6/6	4/6	6.83/3.17	0.59/1.89	0.31	1.60

This report gives the details of the hazard ratio calculation. One line of the report is devoted to each pair of groups.

Groups

These are the two groups reported about on this line, separated by a slash.

Sample Size

These are the sample sizes of the two groups.

Observed Events

These are the number of events (deaths) observed in the two groups.

Expected Events

These are the number of events (deaths) expected in each group under the hypothesis that the two hazard rates are equal.

Hazard Rates

These are the hazard rates of the two groups. The hazard rate is computed as the ratio of the number of observed and expected events.

Cox-Mantel Hazard Ratio

This is the value of the Cox-Mantel hazard ratio. This is the ratio of the two hazard rates.

Hyper. Var.

This is the value of *V*, the hypergeometric variance. This value is used to compute the Mantel-Haenszel hazard ratio and confidence interval.

Hazard Ratio Confidence Interval Section

Groups	Cox-Mantel Hazard Ratio (HR)	Lower 95% C.L. for HR	Upper 95% C.L. for HR	Log Hazard Ratio Value	Log Hazard Ratio S.E.
1/2	3.23	0.85	12.25	1.1733	0.6796
2/1	0.31	0.08	1.17	-1.1733	0.6796

This report gives the details of the Cox-Mantel confidence interval for the hazard ratio. The formulas for these quantities were given earlier in this chapter. One line of the report is devoted to each pair of groups.

Groups

These are the two groups reported about on this line, separated by a slash.

Cox-Mantel Hazard Ratio (HR)

This is the value of the Cox-Mantel hazard ratio.

Lower & Upper 95% C.L. for HR

These are the lower and upper confidence limits of the Cox-Mantel confidence interval of the hazard ratio.

Log Hazard Ratio Value

This is the natural logarithm of the hazard ratio. The logarithmic transformation is applied because the distribution is better approximated by the normal distribution.

Log Hazard Ratio S.E.

This is the standard deviation of the log hazard ratio.

Hazard Ratio Logrank Tests Section

Groups	Cox-Mantel Hazard Ratio	Mantel-Haenszel Hazard Ratio	Cox-Mantel Logrank Test Chi2	Cox-Mantel Prob Level	Mantel-Haenszel Logrank Test Chi2	Mantel-Haenszel Prob Level
1/2	3.23	5.84	3.701	0.0544	4.996	0.0254
2/1	0.31	0.17	3.701	0.0544	4.996	0.0254

This report gives the two logrank tests which test the null hypothesis that the hazard ratio is one (that is, that the hazard rates are equal). The formulas for these quantities were given earlier in this chapter. One line of the report is devoted to each pair of groups.

Groups

These are the two groups reported about on this line, separated by a slash.

Cox-Mantel Hazard Ratio

This is the value of the Cox-Mantel hazard ratio.

Mantel-Haenszel Hazard Ratio

This is the value of the Mantel-Haenszel hazard ratio.

Cox-Mantel Logrank Test

This is the test statistic for the Cox-Mantel logrank test. This value is approximately distributed as a chi-square with one degree of freedom.

Note that this test is more commonly used than the Mantel-Haenszel test.

Cox-Mantel Prob Level

This is the significance level of the Cox-Mantel logrank test. The hypothesis of hazard rate equality is rejected if this value is less than 0.05 (or 0.01).

Mantel-Haenszel Logrank Test

This is the test statistic for the Mantel-Haenszel logrank test. This value is approximately distributed as a chi-square with one degree of freedom.

Mantel-Haenszel Prob Level

This is the significance level of the Mantel-Haenszel logrank test. The hypothesis of hazard rate equality is rejected if this value is less than 0.05 (or 0.01).

Example 3 – Validation of Kaplan-Meier Product Limit Estimator using Collett (1994)

This section presents validation of the Kaplan-Meier product limit estimator and associated statistics. Collett (1994) presents an example on page 5 of the time to discontinuation of use of an IUD. The data are as follows: 10, 13+, 18+, 19, 23+, 30, 36, 38+, 54+, 56+, 59, 75, 93, 97, 104+, 107, 107+, 107+. These data are contained in the COLLETT5.S0 database.

On page 26, Collett (1994) gives the product-limit estimator, its standard deviation, and 95% confidence interval. A partial list of these results is given here:

Time	S(T)	s.e.	95% C.I
10	0.9444	0.0540	(0.839, 1.000)
36	0.7459	0.1170	(0.529, 0.963)
93	0.4662	0.1452	(0.182, 0.751)
107	0.2486	0.1392	(0.000, 0.522)

We will now run these data through this procedure to see that ***NCSS*** gets these same results. You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Kaplan-Meier Curves (Logrank Tests) Fitting window.

1 Open the COLLETT5 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **COLLETT5.S0**.
- Click **Open**.

2 Open the Kaplan-Meier Curves (Logrank Tests) window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Kaplan-Meier Curves (Logrank Tests)**. The Kaplan-Meier Curves (Logrank Tests) procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Variables tab**.
- Set the **Time Variable** to **TIME**.
- Set the **Censor Variable** to **CENSOR**.

4 Specify the reports.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Reports tab**.
- Uncheck all reports except the **Kaplan-Meier Detail** report.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Kaplan-Meier Output

Event Time (T)	Cumulative Survival S(T)	Standard Error of S(T)	Lower 95% C.L. for S(T)	Upper 95% C.L. for S(T)	At Risk	Count	Total Events
10.0	**0.9444**	**0.0540**	**0.8386**	**1.0000**	**18**	**1**	**1**
13.0+					17	1	1
18.0+					16	1	1
19.0	0.8815	0.0790	0.7267	1.0000	15	1	2
23.0+					14	1	2
30.0	0.8137	0.0978	0.6220	1.0000	13	1	3
36.0	**0.7459**	**0.1107**	**0.5290**	**0.9628**	**12**	**1**	**4**
38.0+					11	1	4
54.0+					10	1	4
56.0+					9	1	4
59.0	0.6526	0.1303	0.3972	0.9081	8	1	5
75.0	0.5594	0.1412	0.2827	0.8361	7	1	6
93.0	**0.4662**	**0.1452**	**0.1816**	**0.7508**	**6**	**1**	**7**
97.0	0.3729	0.1430	0.0927	0.6532	5	1	8
104.0+					4	1	8
107.0	**0.2486**	**0.1392**	**0.0000**	**0.5215**	**3**	**1**	**9**
107.0+					2	2	9

You can check this table to see that the results are the same as Collett's.

Example 4 – Validation of Nelson-Aalen Estimator using Klein and Moeschberger (1997)

This section presents validation of the Nelson-Aalen estimator and associated statistics. Klein and Moeschberger (1997) present an example of output for the cumulative hazard function on page 89. The data are available on their website. These data are contained in the BMT.S0 database.

A partial list of these results for the ALL group (our group 1) is given here:

Time	H(T)	s.e.
1	0.0263	0.0263
332	0.5873	0.1449
662	1.0152	0.2185

We will now run these data through this procedure to see that ***NCSS*** gets these same results. You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Kaplan-Meier Curves (Logrank Tests) Fitting window.

1 Open the BMT dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **BMT.S0**.
- Click **Open**.

2 Open the Kaplan-Meier Curves (Logrank Tests) window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Kaplan-Meier Curves (Logrank Tests)**. The Kaplan-Meier Curves (Logrank Tests) procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Variables tab**.
- Set the **Time Variable** to **TIME**.
- Set the **Censor Variable** to **D3**.
- Set the **Group Variable** to **Group**.

4 Specify the reports.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Reports tab**.
- Uncheck all reports except the **Cumulative Hazard Detail** report.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Nelson-Aalen Output

Event Time (T)	Cumulative Hazard H(T)	Standard Error of H(T)	Lower 95% C.L. for H(T)	Upper 95% C.L. for H(T)	At Risk	Count	Total Events
1.0	**0.0263**	**0.0263**	**0.0000**	**0.0779**	**38**	**1**	**1**
55.0	0.0533	0.0377	0.0000	0.1273	37	1	2
74.0	0.0811	0.0468	0.0000	0.1729	36	1	3
86.0	0.1097	0.0549	0.0021	0.2172	35	1	4
104.0	0.1391	0.0623	0.0171	0.2611	34	1	5
107.0	0.1694	0.0692	0.0337	0.3051	33	1	6
109.0	0.2007	0.0760	0.0518	0.3495	32	1	7
110.0	0.2329	0.0825	0.0712	0.3947	31	1	8
122.0	0.2996	0.0950	0.1133	0.4859	30	2	10
129.0	0.3353	0.1015	0.1363	0.5343	28	1	11
172.0	0.3723	0.1081	0.1605	0.5842	27	1	12
192.0	0.4108	0.1147	0.1860	0.6356	26	1	13
194.0	0.4508	0.1215	0.2127	0.6889	25	1	14
226.0+					24	1	14
230.0	0.4943	0.1290	0.2414	0.7472	23	1	15
276.0	0.5397	0.1368	0.2716	0.8079	22	1	16
332.0	**0.5873**	**0.1449**	**0.3034**	**0.8713**	**21**	**1**	**17**
383.0	0.6373	0.1532	0.3370	0.9377	20	1	18
418.0	0.6900	0.1620	0.3724	1.0076	19	1	19
466.0	0.7455	0.1713	0.4098	1.0813	18	1	20
487.0	0.8044	0.1811	0.4494	1.1593	17	1	21
526.0	0.8669	0.1916	0.4913	1.2424	16	1	22
530.0+					15	1	22

609.0	0.9383	0.2045	0.5375	1.3390	14	1	23
662.0	**1.0152**	**0.2185**	**0.5870**	**1.4434**	**13**	**1**	**24**
996.0+					12	1	24
1111.0+					11	1	24
1167.0+					10	1	24
1182.0+					9	1	24
1199.0+					8	1	24
1330.0+					7	1	24
1377.0+					6	1	24
1433.0+					5	1	24
1462.0+					4	1	24
1496.0+					3	1	24
1602.0+					2	1	24
2081.0+					1	1	24

You can check this table to see that the results are the same as Klein and Moeschberger's.

Example 5 – Validation of Logrank Tests using Kleing and Moeschberger (1997)

This section presents validation of the logrank tests. Klein and Moeschberger (1997) present an example of output for the ten logrank tests on page 196. The data are available on their website. These data are contained in the KLEIN6.S0 database.

A list of these results is given here:

Test	Chi-Square	P-Value
Logrank	2.53	0.112
Gehan	0.002	0.964
Tarone-Ware	0.40	0.526
Peto-Peto	1.40	0.237
Modified Peto-Peto	1.28	0.259
Fleming-Harrington(0,1)	9.67	0.002
Fleming-Harrington(1,0)	1.39	0.239
Fleming-Harrington(1,1)	9.83	0.002
Fleming-Harrington(0.5,0.5)	9.28	0.002
Fleming-Harrington(0.5,2)	8.18	0.004

We will now run these data through this procedure to see that ***NCSS*** gets these same results. You may follow along here by making the appropriate entries or load the completed template **Example5** from the Template tab of the Kaplan-Meier Curves (Logrank Tests) Fitting window.

1 Open the KLEIN6 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **KLEIN6.S0**.
- Click **Open**.

2 Open the Kaplan-Meier Curves (Logrank Tests) window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Kaplan-Meier Curves (Logrank Tests).** The Kaplan-Meier Curves (Logrank Tests) procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Variables tab**.
- Set the Time Variable to **TIME**.
- Set the Censor Variable to **CENSOR**.
- Set the Group Variable to **GROUP**.

4 Specify the reports.

- On the Kaplan-Meier Curves (Logrank Tests) window, select the **Reports tab**.
- Uncheck all reports except the **Logrank Test Summary** report.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Logrank Test Output

Test Name	Chi-Square	DF	Prob Level*	Randomization Test Prob Level*	Weighting of Hazard Comparisons Across Time
Logrank	2.530	1	0.1117	0.1210	Equal
Gehan-Wilcoxon	0.002	1	0.9636	0.9670	High++ to Low++
Tarone-Ware	0.403	1	0.5257	0.5480	High to Low
Peto-Peto	1.399	1	0.2369	0.2390	High+ to Low+
Mod. Peto-Peto	1.276	1	0.2587	0.2710	High+ to Low+
F-H (1, 0)	1.387	1	0.2390	0.2450	Almost Equal
F-H (.5, .5)	9.285	1	0.0023	0.0030	Low+ to High+
F-H (1, 1)	9.834	1	0.0017	0.0030	Low to High
F-H (0, 1)	9.668	1	0.0019	0.0020	Low to High
F-H (.5, 2)	8.179	1	0.0042	0.0060	Low++ to High++

You can check this table to see that the results are the same as Klein and Moeschberger's. Note that the order of the tests is different.

Also note that the randomization test probability levels will change slightly from run to run because they are based on Monte Carlo sampling.

Chapter 560

Cumulative Incidence

Introduction

This routine calculates nonparametric, maximum-likelihood estimates and confidence limits of the probability of failure (the *cumulative incidence*) for a particular cause in the presence of other causes. This is sometimes called the problem of *competing risks.*

An often used, though incorrect, approach is to treat all failures from causes other than that of interest as censored observations and estimate the cumulative incidence using 1 - *KM* (Kaplan-Meier estimate). The problem with this approach is that it makes the incorrect assumption that the probability of failing prior to time *t* from other causes is zero. This leads to overestimation of the cumulative incidence. This overestimation can be quite substantial if there are many failures from other causes in the data.

Technical Details

The following results are summarized from Marubini and Valsecchi (1996). Suppose that one of *K* mutually exclusive events may occur to a subject. These events may be failure, death, etc. When an event occurs, the time until it occurred *T* and the type of event *k* is recorded. The experiment may be terminated before any event occurs for some subjects in which case they are called *censored* observations and the time until censoring is recorded. The cause-specific hazard functions are defined as

$$h_k(t) = \lim_{\Delta t \to 0^+} \frac{\Pr\{t \le T < t + \Delta t,\, K = k | T \ge t\}}{\Delta t}, \quad k = 1, 2, ..., K$$

The overall survival function is denoted by

$$S(t) = \Pr(T > t)$$

and the probability of failing for cause *k* is denoted by

$$I_k(t) = \Pr(T \le t, k = K), \quad k = 1, 2, ..., K$$

These cause-specific failure probabilities, also known as the cumulative incidence functions, are defined as

$$I_k(t) = \int_0^t h_k(u) S(u) du, \quad k = 1, 2, ..., K$$

where

$$S(u) = \exp\left\{ -\int_0^u \left(\sum_{k=1}^{k} h_k(y) \right) dy \right\}$$

This makes use of the assumption that a subject must be event free up to time *u* to then fail of cause *k* at time *u*.

If we let d_{kj} denote the number of subjects having event *k* and time t_j, $d_j = \sum_{k=1}^{K} d_{kj}$, and n_j denote the number of subjects at risk at time t_j, the likelihood may be written as

$$L = \prod_{j=1}^{J} \left[\left(\prod_{k=1}^{K} h_{kj}^{d_{kj}} \right) \left(1 - \sum_{k=1}^{K} h_{kj} \right)^{n_j - d_j} \right]$$

The ML estimate of the cause-specific hazard is

$$\hat{h}_{kj} = \frac{d_{kj}}{n_j}$$

and the estimated crude cumulative incidence for event *k* is

$$\hat{I}_k(t) = \sum_{j | t_j < t} \hat{S}(t_{j-1}) \frac{d_{kj}}{n_j}$$

where $\hat{S}(t_{j-1})$ is the usual Kaplan-Meier estimate of survival until time t_{j-1}.

The variance of the crude cumulative incidence is estimated by

$$Var\left[\hat{I}_k(t_j)\right] = \sum_{i=1}^{j} \left\{ \left[\hat{I}_k(t_j) - \hat{I}_k(t_i) \right]^2 \frac{d_i}{n_i(n_i - d_i)} \right\} + \sum_{i=1}^{j} \left[\hat{S}(t_{j-1}) \right]^2 \left(\frac{n_i d_{ki}}{n_i} \right) \left(\frac{d_{ki}}{n_i^2} \right)$$
$$- 2 \sum_{i=1}^{j} \left\{ \left[\hat{I}_k(t_j) - \hat{I}_k(t_i) \right] \hat{S}(t_{j-1}) \frac{d_{ki}}{n_i^2} \right\}$$

Finally, using the above estimate cumulative incidence and its estimated variance, approximate $100(1-\alpha)\%$ confidence intervals may be calculated using

$$\exp\left\{ \log\left[\hat{I}_k(t) \right] \pm z_{1-\alpha/2} \frac{\sqrt{Var\left[\hat{I}_k(t) \right]}}{\hat{I}_k(t)} \right\}$$

This expression guarantees that the resulting values will be between zero and one.

Data Structure

This routine requires at least two variables: one containing the elapsed time values and another containing the type of event. Optional variables include a group identification variable and a frequency variable.

Marubini and Valsecchi (1996) include an example of hypothetical data consisting of two treatment groups and two events. The events are local relapse (1) and distant metastases (2). Censored observations are represented with a zero. Information is available on 35 subjects in each group. These data are stored in the MARUBINI database. The table below shows the data.

MARUBINI dataset

Time	Treatment	Event
1	A	1
13	A	1
17	A	1
30	A	1
34	A	1
41	A	1
78	A	1
100	A	1
119	A	1
169	A	1
1	A	2
6	A	2
8	A	2
13	A	2
13	A	2
15	A	2
33	A	2
37	A	2
44	A	2
45	A	2
63	A	2
80	A	2
89	A	2
89	A	2
91	A	2
132	A	2
144	A	2
171	A	2
183	A	2
240	A	2
34	A	0
60	A	0
63	A	0
149	A	0
207	A	0

Time	Treatment	Event
7	B	1
16	B	1
16	B	1
20	B	1
39	B	1
49	B	1
56	B	1
73	B	1
93	B	1
113	B	1
1	B	2
2	B	2
4	B	2
6	B	2
8	B	2
9	B	2
10	B	2
13	B	2
17	B	2
17	B	2
17	B	2
18	B	2
18	B	2
27	B	2
29	B	2
39	B	2
50	B	2
69	B	2
76	B	2
110	B	2
34	B	0
60	B	0
63	B	0
78	B	0
149	B	0

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Variable

Time Variable

This variable contains the times for each subject. These are elapsed times. If your data are event dates, you must subtract a starting date so that the values are elapsed time. You can scale your data in days, months, or years by dividing by the appropriate constant.

Note that negative time values are treated as missing values. Zero time values are replaced by the value in the Zero Time Replacement option.

Zero Time Replacement

Under normal conditions, a respondent beginning the study is alive and cannot die until after some small period of time has elapsed. Hence, a time value of zero is not defined and is ignored (treated as a missing value). If a zero time value does occur in the database, it is replaced by this positive amount. If you do not want zero time values replaced, enter a “0.0” here.

This option would be used when a “zero” on the database does not actually mean zero time. Instead, it means that the response occurred before the first reading was made and so the actual survival time is only known to be less than one.

Type Variable

Type Variable

Specify the variable containing an identifier for the event type of each observation. If the observation was censored, indicate that using another identifier.

The meaning of these event-type identifiers is designated in the Event Types and Censored Types boxes.

Event Types

This box lists the values of the Type Variable that are to be designated as events. These values may be letters or numbers. For a competing risks analysis, at least two events must be present on the database.

The event may be any occurrence of interest such as failure, death, or recovery. For example, in heart surgery, the events might be death because of heart failure or death for other reasons (accident, cancer, etc.). In this case, two events would be used.

Censor Types

This box lists the values of the Type Variable that are to be designated as being censored. These values may be letters or numbers. Usually, at least one censor value is used.

All of the censor-type values are interpreted as meaning that the observation is right censored. A right censored observation is withdrawn from the study before an event occurs. Hence, you know

that the event will occur after the given length of time, but you do not know when it will occur. For example, the study may end before the patient has died.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Frequency Variable

Frequency Variable

This optional variable gives the count, or frequency, of the time value. Frequency values must be positive integers. When omitted, each row has a frequency of one.

This variable is often used to indicate the number of CENSORED values at the end of a study. It may also be used to indicate TIES for event-type data.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report - Cumulative Survival Summary Report

These options specify whether to display the corresponding reports.

Report Options

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Events to Report

Events Reported On

This option limits the reports to those event-types that are of primary interest. Separate sets of reports and plots are generated for each event type listed here. Enter ALL to have all event types reported on.

Note that this option does not change the analysis, just the reports.

Report Options – Times Displayed on Summary Reports

Summary Report Times

Specify a list of times that are reported on in the summary reports. The regular reports contain output for each time on the database. These may not represent the most useful time values. For example, in a study that lasts two years, you may want a summary report for every six months.

To specify the time values, use numbers separated by commas or blanks. You may specify a sequence of values with a colon, putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers.

Use (10) alone to specify ten, equal-spaced values between zero and the maximum (zero not included).

Report Options – Decimal Places

Time

This option specifies the number of decimal places shown on reported time values.

Probability

This option specifies the number of decimal places shown on reported probability and hazard values.

Report Options – Title

Report Title

This option specifies a title to appear at the top of each page.

Plots Tab

The following options control which plots are displayed and the format of those plots.

Select Plots

Individual Incidence Plots - Combined Survival Plots

Specify whether to display each of these plots.

Plot Options

These options control the contents and arrangement of the plots.

Plot Options – Plot Arrangement

Two Plots Per Line

Specify whether to display one or two plots per line. Choosing two plots forces the plots to be smaller so that two will fit across a page.

Plot Options – Plot Contents

Cumulative Line

Specify whether to display the estimated line (cumulative incidence or cumulative survival) on the plots.

Confidence Limits on Individual & Combined Plots

Specify whether to display the confidence limits on each type of the plot.

Legend on Individual Plots & Combined Plots

Specify whether to display the legend on each of the plots.

Legend Text

Specify the text that is to be displayed as the heading of the legend. When {G} is entered, it is automatically replaced by the name of the group variable.

Censor Tickmarks

This option indicates the size of the tickmarks (if any) showing where the censored points fall on the curve. The values are at a scale of 1000 equals one inch. Enter 0 for no censor tickmarks or 100 to display tickmarks.

Cum Inc Plot and Cum Surv Plot Tabs

These options control the attributes of the cumulative incidence and survival curves.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Title Lines 1 and 2

These options contain the text of the titles. The characters *{Y}*, *{X}*, and *{M}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines Tab

These options specify the attributes of the lines used for each group in the various curves.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Storage Tab

These options control the storage of information back to the database for further use.

Data Storage Variables

Group - Product Limit

Each of these options let you specify columns (variables) on the database to which the corresponding data are automatically stored. Warning: existing data are replaced, so make sure that the columns you select are empty.

Note that no attempt is made to store the time values in their original order. That’s why you have to store the Group and Event Type to identify the incidence values.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Cumulative Incidence

This section presents an example of how to generate cumulative incidence reports. The data used were shown above and are found in the MARUBINI database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Cumulative Incidence window.

1 Open the MARUBINI dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MARUBINI.s0**.
- Click **Open**.

2 Open the Cumulative Incidence window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Cumulative Incidence**. The Cumulative Incidence procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Cumulative Incidence window, select the **Variables tab**.
- Enter **1** in the **Time Variable** box. This sets the Time Variable to the first variable on the database which is the Time variable.
- Enter **3** in the **Type Variable** box.
- Enter **1 2** in the **Event Types** box.
- Enter **0** in the **Censor Types** box.
- Enter **2** in the **Group Variable** box.

4 Specify the reports.

- On the Cumulative Incidence window, select the **Reports tab**.
- Enter **1** in the **Events Reported On** box.
- Enter **25 50 75 100 125 150 175 200** in the **Summary Report Times** box.

5 Specify the plots.

- On the Cumulative Incidence window, select the **Plots tab**.
- Check the **Two Plots Per Line** box.
- Check the **Confidence Limits on Individual Plots** box.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section. Treatment = A

Type of Observation	Values	Count	Minimum	Maximum
Censored	5	5	34	207
Treatment = 1	10	10	1	169
Treatment = 2	18	20	1	240
Total	33	35	0	240

This section of the report displays information about the database. It is especially useful to allow you to check for obvious data-entry errors. The Values column gives the number of unique time values. The Count column uses the values in the Frequency Variable when it was used.

Cumulative Incidence Detail Report

Cumulative Incidence Detail Report for Event = 1 and Treatment = A

Time	Number At Risk	Events of Type 1	Events of All Types	Cumulative Incidence	Lower 95% C.L. Cum. Inc.	Upper 95% C.L. Cum. Inc.	Standard Error of Cum. Inc.	1 - Product Limit
1.0	35	1	2	0.0286	0.0041	0.1972	0.0282	0.0571
6.0	33	0	1	0.0286	0.0041	0.1972	0.0282	0.0857
8.0	32	0	1	0.0286	0.0041	0.1972	0.0282	0.1143
13.0	31	1	3	0.0571	0.0149	0.2195	0.0392	0.2000
15.0	28	0	1	0.0571	0.0149	0.2195	0.0392	0.2286
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
240.0	1	0	1	0.3144	0.1871	0.5285	0.0833	1.0000

This report displays the cumulative incidence values along with their confidence intervals and standard errors.

Time

This is the time value, t_j, being reported on. These values are from the dataset being analyzed. Note that tied values are combined.

Number at Risk

This is the number of individuals at risk, n_j, just before time t_j.

Events of Type 1

This is the number of events of the type indicated (in this report, the type is 1), d_{kj}, that occurred at t_j.

Events of All Types

This is the number of events of all types, d_{kj}, that occurred at time t_j. Note that censored observations have a zero in this column.

Cumulative Incidence

This is the cumulative incidence, $\hat{I}_k\left(t_j\right)$. This is the cumulative probability of event *k* up through the current time value, accounting for all other events. The formula used to calculate this value was given in the Technical Details section earlier in this chapter.

Lower and Upper 95% C.L. Cum. Inc.

These are confidence limits for the above cumulative incidence.

Standard Error of Cum. Inc.

This is the estimated standard error of the cumulative incidence. It is calculated using $\sqrt{Var\left[\hat{I}_k\left(t_j\right)\right]}$.

1 – Product Limit

This is one minus the Kaplan-Meier product limit estimate. This is a cumulative incidence measure calculated assuming that there are no other possible events other than the event of interest.

Cumulative Survival Detail Report

Cumulative Survival Detail Report for Event = 1 and Treatment = A

Time	Number At Risk	Events of Type 1	Events of All Types	Cumulative Survival	Lower 95% C.L. Cum. Surv.	Upper 95% C.L. Cum. Surv.	Standard Error of Cum. Surv.	Product Limit
1.0	35	1	2	0.9714	0.9959	0.8028	0.0282	0.9429
6.0	33	0	1	0.9714	0.9959	0.8028	0.0282	0.9143
8.0	32	0	1	0.9714	0.9959	0.8028	0.0282	0.8857
13.0	31	1	3	0.9429	0.9851	0.7805	0.0392	0.8000
15.0	28	0	1	0.9429	0.9851	0.7805	0.0392	0.7714
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
240.0	1	0	1	0.6856	0.8129	0.4715	0.0833	0.0000

This report displays the cumulative survival values along with their confidence intervals and standard errors. The cumulative survival values are equal to one minus the cumulative incidence values.

Cumulative Incidence Summary Report

Cumulative Incidence Summary Report for Event = 1 and Treatment = A

Time	Number At Risk	Events of Type 1	Events of All Types	Cumulative Incidence	Lower 95% C.L. Cum. Inc.	Upper 95% C.L. Cum. Inc.	Standard Error of Cum. Inc.	1 - Product Limit
25.0	27	1	1	0.0857	0.0290	0.2529	0.0473	0.2571
50.0	19	0	1	0.1727	0.0834	0.3578	0.0642	0.4623
75.0	17	0	1	0.1727	0.0834	0.3578	0.0642	0.4940
100.0	10	1	1	0.2402	0.1308	0.4411	0.0745	0.6964
125.0	9	1	1	0.2739	0.1564	0.4797	0.0783	0.7301
150.0	6	0	0	0.2739	0.1564	0.4797	0.0783	0.7976
175.0	4	0	1	0.3144	0.1871	0.5285	0.0833	0.8786
200.0	3	0	1	0.3144	0.1871	0.5285	0.0833	0.9190

This report displays the cumulative incidence values at designated time values. All definitions are the same as in the Cumulative Incidence report. The number at risk and the numbers of events are for the last actual time value in the data before the current time value. For example, the values shown for the time of 75.0 are actually the values for the time of 63.0 on the database since 63.0 is the largest time value before 75.0.

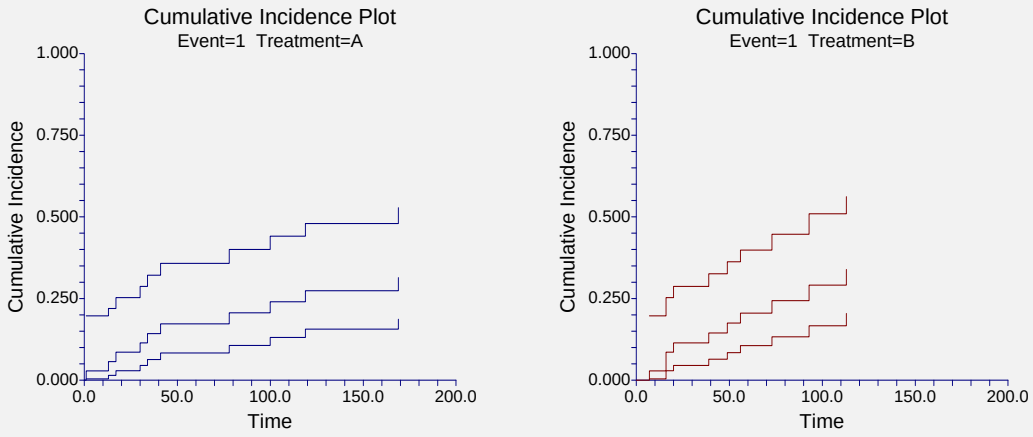
Cumulative Survival Summary Report

Cumulative Incidence Summary Report for Event = 1 and Treatment = A

Time	Number At Risk	Events of Type 1	Events of All Types	Cumulative Incidence	Lower 95% C.L. Cum. Inc.	Upper 95% C.L. Cum. Inc.	Standard Error of Cum. Inc.	1 - Product Limit
25.0	27	1	1	0.9143	0.9710	0.7471	0.0473	0.7429
50.0	19	0	1	0.8273	0.9166	0.6422	0.0642	0.5377
75.0	17	0	1	0.8273	0.9166	0.6422	0.0642	0.5060
100.0	10	1	1	0.7598	0.8692	0.5589	0.0745	0.3036
125.0	9	1	1	0.7261	0.8436	0.5203	0.0783	0.2699
150.0	6	0	0	0.7261	0.8436	0.5203	0.0783	0.2024
175.0	4	0	1	0.6856	0.8129	0.4715	0.0833	0.1214
200.0	3	0	1	0.6856	0.8129	0.4715	0.0833	0.0810

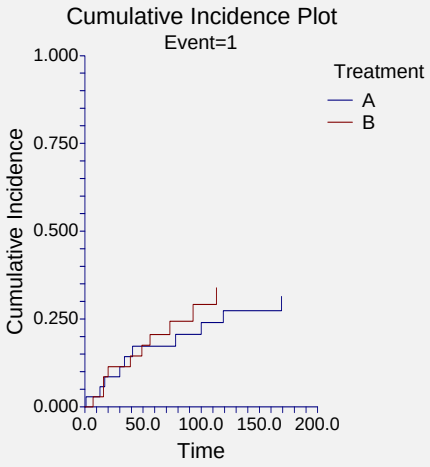
This report displays the survival values at designated time values. All definitions are the same as in the Cumulative Survival report. The number at risk and the numbers of events are for the last actual time value in the data before the current time value. For example, the values shown for the time of 75.0 are actually the values for the time of 63.0 on the database since 63.0 is the largest time value before 75.0.

Individual Cumulative Incidence Plots



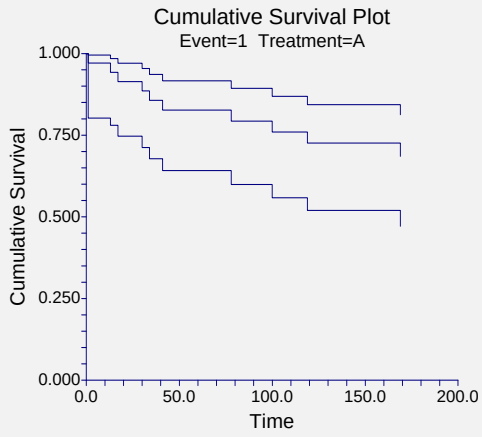
These plots show the cumulative incidence curve surrounded by the 95% confidence interval.

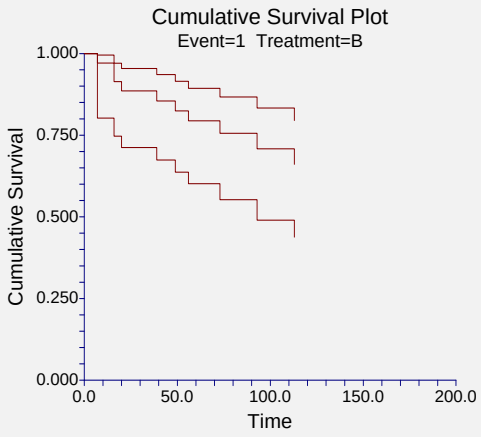
Combined Cumulative Incidence Plot



This plot shows the cumulative incidence for each of the two groups.

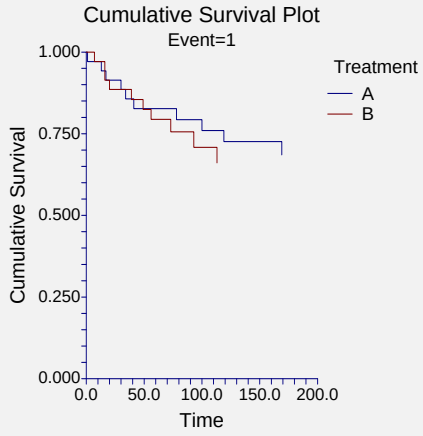
Individual Cumulative Survival Plots





These plots display the cumulative survival values and confidence intervals for each group, one group per plot.

Combined Cumulative Survival Plot



This plot displays the cumulative survival values for both groups.

Validation of Cumulative Incidence Values using Marubini and Valsecchi (1996)

Marubini and Valsecchi (1996) reported cumulative incidence values calculated for the data in the MARUBINI database. For group B, for the time value of 16, they calculated a cumulative incidence of 0.08571, a standard error of 0.04732, and confidence limits of 0.02905 and 0.25292. You can check the ***NCSS*** output to see that it obtains these same values, which validates its accuracy.

Chapter 565

Cox Regression

Introduction

This program performs Cox (proportional hazards) regression analysis, which models the relationship between a set of one or more covariates and the hazard rate. Covariates may be discrete or continuous. Cox's proportional hazards regression model is solved using the method of marginal likelihood outlined in Kalbfleisch (1980).

This routine can be used to study the impact of various factors on survival. You may be interested in the impact of diet, age, amount of exercise, and amount of sleep on the survival time after an individual has been diagnosed with a certain disease such as cancer. Under normal conditions, the obvious statistical tool to study the relationship between a response variable (survival time) and several explanatory variables would be multiple regression. Unfortunately, because of the special nature of survival data, multiple regression is not appropriate. Survival data usually contain censored data and the distribution of survival times is often highly skewed. These two problems invalidate the use of multiple regression. Many alternative regression methods have been suggested. The most popular method is the proportional hazard regression method developed by Cox (1972). Another method, Weibull regression, is available in ***NCSS*** in the Distribution Regression procedure.

Further Reading

Several books provide in depth coverage of Cox regression. These books assume a familiarity with basic statistical theory, especially with regression analysis. Collett (1994) provides a comprehensive introduction to the subject. Hosmer and Lemeshow (1999) is almost completely devoted to this subject. Therneau and Grambsch (2000) provide a complete and up-to-date discussion of this subject. We found their discussion of residual analysis very useful. Klein and Moeschberger (1997) provides a very readable account of survival analysis in general and includes a lucid account of Cox regression.

The Cox Regression Model

Survival analysis refers to the analysis of elapsed time. The response variable is the time between a *time origin* and an *end point.* The end point is either the occurrence of the event of interest, referred to as a *death* or *failure*, or the end of the subject's participation in the study. These elapsed times have two properties that invalidate standard statistical techniques, such as t-tests, analysis of variance, and multiple regression. First of all, the time values are often positively skewed. Standard statistical techniques require that the data be normally distributed. Although this skewness could be corrected with a transformation, it is easier to adopt a more realistic data distribution.

The second problem with survival data is that part of the data are *censored*. An observation is censored when the end point has not been reached when the subject is removed from study. This may be because the study ended before the subject's response occurred, or because the subject withdrew from active participation. This may be because the subject died for another reason, because the subject moved, or because the subject quit following the study protocol. All that is known is that the response of interest did not occur while the subject was being studied.

When analyzing survival data, two functions are of fundamental interest—the *survivor function* and the *hazard function*. Let *T* be the survival time. That is, *T* is the elapsed time from the beginning point, such as diagnosis of cancer, and death due to that disease. The values of *T* can be thought of as having a *probability distribution*. Suppose the *probability density function* of the random variable *T* is given by $f(T)$. The *probability distribution function* of *T* is then given by

$$\begin{aligned} F(T) &= \Pr(t < T) \\ &= \int_0^T f(t)\mathrm{d}t \end{aligned}$$

The *survivor function*, $S(T)$, is the probability that an individual survives past *T*. This leads to

$$\begin{aligned} S(T) &= \Pr(T \geq t) \\ &= 1 - F(T) \end{aligned}$$

The *hazard function* is the probability that a subject experiences the event of interest (death, relapse, etc.) during a small time interval given that the individual has survived up to the beginning of that interval. The mathematical expression for the hazard function is

$$\begin{aligned} h(T) &= \lim_{\Delta T \to 0} \frac{\Pr(T \leq t < (T + \Delta T) \mid T \leq t)}{\Delta T} \\ &= \lim_{\Delta T \to 0} \frac{F(T + \Delta T) - F(T)}{\Delta T} \\ &= \frac{f(T)}{S(T)} \end{aligned}$$

The cumulative hazard function $H(T)$ is the sum of the individual hazard rates from time zero to time *T*. The formula for the cumulative hazard function is

$$H(T) = \int_0^T h(u)\mathrm{d}u$$

Thus, the hazard function is the derivative, or slope, of the cumulative hazard function. The cumulative hazard function is related to the cumulative survival function by the expression

$$S(T) = e^{-H(T)}$$

or

$$H(T) = -\ln(S(T))$$

We see that the distribution function, the hazard function, and the survival function are mathematically related. As a matter of convenience and practicality, the hazard function is used in the basic regression model.

Cox (1972) expressed the relationship between the hazard rate and a set of covariates using the model

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^{p} x_i \beta_i$$

or

$$h(T) = h_0(T) e^{\sum_{i=1}^{p} x_i \beta_i}$$

where $x_1, x_2, \cdots, x_p$ are covariates, $\beta_1, \beta_2, \cdots, \beta_p$ are regression coefficients to be estimated, T is the elapsed time, and $h_0(T)$ is the baseline hazard rate when all covariates are equal to zero. Thus the linear form of the regression model is

$$\ln\left[\frac{h(T)}{h_0(T)}\right] = \sum_{i=1}^{p} x_i \beta_i$$

Taking the exponential of both sides of the above equation, we see that this is the ratio between the actual hazard rate and the baseline hazard rate, sometimes called the *relative risk*. This can be rearranged to give the model

$$\begin{aligned}\frac{h(T)}{h_0(T)} &= \exp\left(\sum_{i=1}^{p} x_i \beta_i\right) \\ &= e^{x_1\beta_1} e^{x_2\beta_2} \cdots e^{x_p\beta_p}\end{aligned}$$

The regression coefficients can thus be interpreted as the relative risk when the value of the covariate is increased by one unit.

Note that unlike most regression models, this model does not include an intercept term. This is because if an intercept term were included, it would become part of $h_0(T)$.

Also note that the above model does not include T on the right-hand side. That is, the relative risk is constant for all time values. This is why the method is called *proportional hazards*.

An interesting attribute of this model is that you only need to use the ranks of the failure times to estimate the regression coefficients. The actual failure times are not used except to generate the ranks. Thus, you will achieve the same regression coefficient estimates regardless of whether you enter the time values in days, months, or years.

Cumulative Hazard

Under the proportional hazards regression model, the cumulative hazard is

$$\begin{aligned}H(T, X) &= \int_0^T h(u, X)\, du \\ &= \int_0^T h_0(u) e^{\sum_{i=1}^{p} x_i \beta_i}\, du\end{aligned}$$

$$= e^{\sum_{i=1}^{p} x_i\beta_i} \int_0^T h_0(u)du$$
$$= H_0(T)e^{\sum_{i=1}^{p} x_i\beta_i}$$

Note that the survival time *T* is present in $H_0(T)$, but not in $e^{\sum_{i=1}^{p} x_i\beta_i}$. Hence, the cumulative hazard up to time *T* is represented in this model by a baseline cumulative hazard $H_0(T)$ which is adjusted by the covariates by multiplying by the factor $e^{\sum_{i=1}^{p} x_i\beta_i}$.

Cumulative Survival

Under the proportional hazards regression model, the cumulative survival is

$$\begin{aligned} S(T,X) &= \exp\left(-H(T,X)\right) \\ &= \exp\left(-H_0(T)e^{\sum_{i=1}^{p} x_i\beta_i}\right) \\ &= \left[e^{-H_0(T)}\right]^{e^{\sum_{i=1}^{p} x_i\beta_i}} \\ &= S_0(T)^{e^{\sum_{i=1}^{p} x_i\beta_i}} \end{aligned}$$

Note that the survival time *T* is present in $S_0(T)$, but not in $e^{\sum_{i=1}^{p} x_i\beta_i}$.

A Note On Using e

The discussion that follows uses the terms exp(*x*) and e^x. These terms are identical. That is

$$\begin{aligned} \exp(x) &= e^x \\ &= (2.71828182846)^x \end{aligned}$$

The decision as to which form to use depends on the context. The preferred form is e^x. But often, the expression used for *x* becomes so small that it cannot be printed. In these situations, the exp(*x*) form will be used.

One other point needs to be made while we are on this subject. People often wonder why we use the number *e*. After all, *e* is an unfamiliar number that cannot be expressed exactly. Why not use a more common number like 2, 3, or 10? The answer is that it does matter because the choice of the base is arbitrary in that you can easily switch from one base to another. That is, it is easy to find constants *a*, *b*, and *c* so that

$$e = 2^a = 3^b = 10^c$$

In fact, *a* is 1/ln(2) = 1.4427, *b* is 1/ln(3)=0.9102, and *c* is 1/ln(10) = 0.4343. Using these constants, it is easy to switch from one base to another. For example, suppose a calculate only computes 10^x and we need the value of e^3. This can be computed as follows

$$\begin{aligned} e^3 &= \left(10^{0.4343}\right)^3 \\ &= 10^{3(0.4343)} \\ &= 10^{1.3029} \\ &= 20.0855 \end{aligned}$$

The point is, it is simple to change from base *e* to base 3 to base 10. The number *e* is used for mathematical convenience.

Maximum Likelihood Estimation

Let $t = 1, \cdots, M$ index the *M* unique failure times $T_1, T_2, ..., T_M$. Note that *M* does not include duplicate times or censored observations. The set of all failures (deaths) that occur at time T_t is referred to as D_t. Let c and $d = 1, \cdots, m_t$ index the members of D_t. The set of all individuals that are at risk immediately before time T_t is referred to as R_t. This set, often called the *risk set*, includes all individuals that fail at time T_t as well as those that are censored or fail at a time later than T_t. Let $r = 1, \cdots, n_t$ index the members of R_t. Let *X* refer to a set of *p* covariates. These covariates are indexed by the subscripts *i*, *j*, or *k*. The values of the covariates at a particular failure time T_d are written $x_{1d}, x_{2d}, \cdots, x_{pd}$ or x_{id} in general. The regression coefficients to be estimated are $\beta_1, \beta_2, \cdots, \beta_p$.

The Log Likelihood

When there are no ties among the failure times, the log likelihood is given by Kalbfleisch and Prentice (1980) as

$$\begin{aligned} LL(\beta) &= \sum_{t=1}^{M} \left\{ \left(\sum_{i=1}^{p} x_{it}\beta_i \right) - \ln\left(\sum_{r \in R_t} \exp\left(\sum_{i=1}^{p} x_{ir}\beta_i \right) \right) \right\} \\ &= \sum_{t=1}^{M} \left\{ \sum_{i=1}^{p} x_{it}\beta_i - \ln\left(G_{R_t}\right) \right\} \end{aligned}$$

where

$$G_R = \sum_{r \in R} \exp\left(\sum_{i=1}^{p} x_{ir}\beta_i \right)$$

The following notation for the first-order and second-order partial derivatives will be useful in the derivations in this section.

$$\begin{aligned} H_{jR} &= \frac{\partial G_R}{\partial \beta_j} \\ &= \sum_{r \in R} x_{jr} \exp\left(\sum_{i=1}^{p} x_{ir}\beta_i \right) \end{aligned}$$

$$A_{jkR} = \frac{\partial^2 G_R}{\partial\beta_j \partial\beta_k} = \frac{\partial H_{jR}}{\partial\beta_k} = \sum_{r\in R} x_{jr} x_{kr} \exp\left(\sum_{i=1}^{p} x_{ir}\beta_i\right)$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first and second order partial derivatives. The first order partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial\beta_j} = \sum_{t=1}^{M}\left\{x_{jt} - \frac{H_{jR_t}}{G_{R_t}}\right\}$$

The second order partial derivatives, which are the information matrix, are

$$I_{jk} = \sum_{t=1}^{M} \frac{1}{G_{R_t}}\left(A_{jkR_t} - \frac{H_{jR_t} H_{kR_t}}{G_{R_t}}\right)$$

When there are failure time ties (note that censor ties are not a problem), the exact likelihood is very cumbersome. ***NCSS*** allows you to select either the approximation proposed by Breslow (1974) or the approximation given by Efron (1977). Breslow's approximation was used by the first Cox regression programs, but Efron's approximation provides results that are usually closer to the results given by the exact algorithm and it is now the preferred approximation (see for example Homer and Lemeshow (1999). We have included Breslow's method because of its popularity. For example, Breslow's method is the default method used in SAS.

Breslow's Approximation to the Log Likelihood

The log likelihood of Breslow's approximation is given by Kalbfleisch and Prentice (1980) as

$$LL(\beta) = \sum_{t=1}^{M}\left\{\left(\sum_{d\in D_t}\sum_{i=1}^{p} x_{id}\beta_i\right) - m_t \ln\left[\sum_{r\in R_t}\exp\left(\sum_{i=1}^{p} x_{ir}\beta_i\right)\right]\right\} = \sum_{t=1}^{M}\left\{\sum_{d\in D_t}\sum_{i=1}^{p} x_{id}\beta_i - m_t \ln\left(G_{R_t}\right)\right\}$$

where

$$G_R = \sum_{r\in R}\exp\left(\sum_{i=1}^{p} x_{ir}\beta_i\right)$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first-order and second-order partial derivatives. The first order partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{t=1}^{M} \left\{ \left(\sum_{d \in D_t} x_{jd} \right) - \left(m_t \frac{H_{jR_t}}{G_{R_t}} \right) \right\}$$

The negative of the second-order partial derivatives, which form the information matrix, are

$$I_{jk} = \sum_{t=1}^{M} \frac{m_t}{G_{R_t}} \left(A_{jkR_t} - \frac{H_{jR_t} H_{kR_t}}{G_{R_t}} \right)$$

Efron's Approximation to the Log Likelihood

The log likelihood of Efron's approximation is given by Kalbfleisch and Prentice (1980) as

$$LL(\beta) = \sum_{t=1}^{M} \left\{ \sum_{d \in D_t} \sum_{i=1}^{p} x_{id} \beta_i - \sum_{d \in D_t} \ln \left[\sum_{r \in R_t} \exp\left(\sum_{i=1}^{p} x_{ir} \beta_i \right) - \frac{d-1}{m_t} \sum_{c \in D_t} \exp\left(\sum_{i=1}^{p} x_{ic} \beta_i \right) \right] \right\}$$
$$= \sum_{t=1}^{M} \left\{ \sum_{d \in D_t} \sum_{i=1}^{p} x_{id} \beta_i - \sum_{d \in D_t} \ln \left[G_{R_t} - \frac{d-1}{m_t} G_{D_t} \right] \right\}$$

The maximum likelihood solution is found by the Newton-Raphson method. This method requires the first and second order partial derivatives. The first partial derivatives are

$$U_j = \frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{t=1}^{M} \sum_{d \in D_t} \left(x_{jd} - \frac{H_{jR_t} - \left(\frac{d-1}{m_t} \right) H_{jD_t}}{G_{R_t} - \left(\frac{d-1}{m_t} \right) G_{D_t}} \right)$$
$$= \sum_{t=1}^{M} \sum_{d \in D_t} x_{jd} - \sum_{t=1}^{M} \sum_{d=1}^{m_t} \left(\frac{H_{jR_t} - \left(\frac{d-1}{m_t} \right) H_{jD_t}}{G_{R_t} - \left(\frac{d-1}{m_t} \right) G_{D_t}} \right)$$

The second partial derivatives provide the information matrix which estimates the covariance matrix of the estimated regression coefficients. The negative of the second partial derivatives are

$$I_{jk} = -\frac{\partial^2 LL(\beta)}{\partial \beta_j \partial \beta_k} = \sum_{t=1}^{M} \sum_{d=1}^{m_t} \frac{\left(G_{R_t} - \left(\frac{d-1}{m_t} \right) G_{D_t} \right) \left(A_{jkR_t} - \left(\frac{d-1}{m_t} \right) A_{jkD_t} \right) - \left(H_{jR_t} - \left(\frac{d-1}{m_t} \right) H_{jD_t} \right) \left(H_{kR_t} - \left(\frac{d-1}{m_t} \right) H_{kD_t} \right)}{\left(G_{R_t} - \left(\frac{d-1}{m_t} \right) G_{D_t} \right)^2}$$

Estimation of the Survival Function

Once the maximum likelihood estimates have been obtained, it may be of interest to estimate the survival probability of a new or existing individual with specific covariate settings at a particular point in time. The methods proposed by Kalbfleisch and Prentice (1980) are used to estimate the survival probabilities.

Cumulative Survival

This estimates the cumulative survival of an individual with a set of covariates all equal to zero. The survival for an individual with covariate values of X_0 is

$$\begin{aligned} S(T \mid X_0) &= \exp(H(T \mid X_0)) \\ &= \exp\left(H_0(T \mid X_0) \exp \sum_{i=1}^{p} x_{i0}\beta \right) \\ &= \left[S_0(T)\right]^{\exp \sum_{i=1}^{p} x_{i0}\beta} \end{aligned}$$

The estimate of the baseline survival function $S_0(T)$ is calculated from the cumulated hazard function using

$$S_0(T_0) = \prod_{T_t \le T_0} \alpha_t$$

where

$$\begin{aligned} \alpha_t &= \frac{S(T_t)}{S(T_{t-1})} \\ &= \left[\frac{S_0(T_t)}{S_0(T_{t-1})} \right]^{\exp\left(\sum_{i=1}^{p} x_{it}\beta_i \right)} \\ &= \left[\frac{S_0(T_t)}{S_0(T_{t-1})} \right]^{\theta_t} \end{aligned}$$

where

$$\theta_r = \exp\left(\sum_{i=1}^{p} x_{ir}\beta_i \right)$$

The value of α_t, the conditional baseline survival probability at time *T*, is the solution to the conditional likelihood equation

$$\sum_{d \in D_t} \frac{\theta_d}{1 - \alpha_t^{\theta_d}} = \sum_{r \in R_t} \theta_r$$

When there are no ties at a particular time point, D_t contains one individual and the above equation can be solved directly, resulting in the solution

$$\hat{\alpha}_t = \left[1 - \frac{\hat{\theta}_t}{\sum_{r \in R_t} \hat{\theta}_r} \right]^{\hat{\theta}_t^{-1}}$$

When there are ties, the equation must be solved iteratively. The starting value of this iterative process is

$$\hat{\alpha}_t = \exp\left(\frac{-m_t}{\sum_{r \in R_t} \hat{\theta}_r}\right)$$

Baseline Hazard Rate

Hosmer and Lemeshow (1999) estimate the baseline hazard rate $h_0(T_t)$ as follows

$$h_0(T_t) = 1 - \alpha_t$$

They mention that this estimator will typically be too unstable to be of much use. To overcome this, you might smooth these quantities using lowess function of the Scatter Plot program.

Cumulative Hazard

An estimate of the cumulative hazard function $H_0(T)$ derived from relationship between the cumulative hazard and the cumulative survival. The estimated baseline survival is

$$\hat{H}_0(T) = -\ln\left(\hat{S}_0(T)\right)$$

This leads to the estimated cumulative hazard function is

$$\hat{H}(T) = -\exp\left(\sum_{i=1}^{p} x_i \hat{\beta}_i\right) \ln\left(\hat{S}_0(T)\right)$$

Cumulative Survival

The estimate of the cumulative survival of an individual with a set of covariates values of X_0 is

$$\hat{S}(T \mid X_0) = \hat{S}_0(T)^{\exp\sum_{i=1}^{p} x_{i0}\hat{\beta}_i}$$

Statistical Tests and Confidence Intervals

Inferences about one or more regression coefficients are all of interest. These inference procedures can be treated by considering hypothesis tests and/or confidence intervals. The inference procedures in Cox regression rely on large sample sizes for accuracy.

Two tests are available for testing the significance of one or more independent variables in a regression: the likelihood ratio test and the Wald test. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation.

These two testing procedures will be described next.

Likelihood Ratio and Deviance

The *Likelihood Ratio* test statistic is -2 times the difference between the log likelihoods of two models, one of which is a subset of the other. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF)

of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. The test is named as a ratio rather than a difference since the difference between two log likelihoods is equal to the log of the ratio of the two likelihoods. That is, if L_{full} is the log likelihood of the full model and L_{subset} is the log likelihood of a subset of the full model, the likelihood ratio is defined as

$$\begin{aligned} LR &= -2\left[L_{\text{subset}} - L_{\text{full}}\right] \\ &= -2\left[\ln\left(\frac{l_{\text{subset}}}{l_{\text{full}}}\right)\right] \end{aligned}$$

Note that the -2 adjusts *LR* so the chi-square distribution can be used to approximate its distribution.

The likelihood ratio test is the test of choice in Cox regression. Various simulation studies have shown that it is more accurate than the Wald test in situations with small to moderate sample sizes. In large samples, it performs about the same. Unfortunately, the likelihood ratio test requires more calculations than the Wald test, since it requires the fitting of two maximum-likelihood models.

Deviance

When the full model in the likelihood ratio test statistic is the saturated model, *LR* is referred to as the *deviance*. A saturated model is one which includes all possible terms (including interactions) so that the predicted values from the model equal the original data. The formula for the deviance is

$$D = -2\left[L_{\text{Reduced}} - L_{\text{Saturated}}\right]$$

The deviance in Cox regression is analogous to the residual sum of squares in multiple regression. In fact, when the deviance is calculated in multiple regression, it is equal to the sum of the squared residuals.

The change in deviance, ΔD, due to excluding (or including) one or more variables is used in Cox regression just as the partial *F* test is used in multiple regression. Many texts use the letter *G* to represent ΔD. Instead of using the *F* distribution, the distribution of the change in deviance is approximated by the chi-square distribution. Note that since the log likelihood for the saturated model is common to both deviance values, ΔD can be calculated without actually fitting the saturated model. This fact becomes very important during subset selection. The formula for ΔD for testing the significance of the regression coefficient(s) associated with the independent variable *X*1 is

$$\begin{aligned} \Delta D_{X1} &= D_{\text{without } X1} - D_{\text{with } X1} \\ &= -2\left[L_{\text{without } X1} - L_{\text{Saturated}}\right] + 2\left[L_{\text{with } X1} - L_{\text{Saturated}}\right] \\ &= -2\left[L_{\text{without } X1} - L_{\text{with } X1}\right] \end{aligned}$$

Note that this formula looks identical to the likelihood ratio statistic. Because of the similarity between the change in deviance test and the likelihood ratio test, their names are often used interchangeably.

Wald Test

The Wald test will be familiar to those who use multiple regression. In multiple regression, the common *t*-test for testing the significance of a particular regression coefficient is a Wald test. In

Cox regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$z_j = \frac{b_j}{s_{b_j}}$$

where s_{b_j} is an estimate of the standard error of b_j provided by the square root of the corresponding diagonal element of the covariance matrix, $V\left(\hat{\beta}\right) = I^{-1}$.

With large sample sizes, the distribution of z_j is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as 'adequate.'

The Wald test is used in ***NCSS*** to test the statistical significance of individual regression coefficients.

Confidence Intervals

Confidence intervals for the regression coefficients are based on the Wald statistics. The formula for the limits of a $100(1-\alpha)\%$ two-sided confidence interval is

$$b_j \pm |z_{\alpha/2}| s_{b_j}$$

R-Squared

Hosmer and Lemeshow (1999) indicate that at the time of the writing of their book, there is no single, easy to interpret measure in Cox regression that is analogous to R^2 in multiple regression. They indicate that if such a measure "must be calculated" they would use

$$R_p^2 = 1 - \exp\left[\frac{2}{n}\left(L_0 - L_p\right)\right]$$

where L_0 is the log likelihood of the model with no covariates, *n* is the number of observations (censored or not), and L_p is the log likelihood of the model that includes the covariates.

Subset Selection

Subset selection refers to the task of finding a small subset of the available regressor variables that does a good job of predicting the dependent variable. Because Cox regression must be solved iteratively, the task of finding the best subset can be time consuming. Hence, techniques which look at all possible combinations of the regressor variables are not feasible. Instead, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Before discussing the details of these two algorithms, it is important to comment on a couple of issues that can come up. The first issue is what to do about the binary variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms used here search on model terms rather than on the individual variables. Thus, the whole set of binary variables associated with a given term are considered together for inclusion in, or deletion from, the model. Its all or none. Because of the

time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and designate them as Numeric Variables.

Hierarchical Models

A second issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to only consider hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of *R*-squared. Enter this term into the model.
3. Continue adding terms until a preset limit on the maximum number of terms in the model is reached.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations so that other, more time consuming methods, are not feasible, or when you have far too many possible regressor variables and you want to reduce the number of terms in the selection pool.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of *R*-squared. If a switch can be found, it is made and the candidate terms are again searched to determine if another switch can be made.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some F-to-enter and F-to-remove tests whose properties are not well understood to begin with.

Residuals

The following presentation summarizes the discussion on residuals found in Klein and Moeschberger (1997) and Hosmer and Lemeshow (1999). For a more thorough treatment of this topic, we refer you to either of these books.

In most settings in which residuals are studied, the dependent variable is predicted using a model based on the independent variables. In these cases, the residual is simply the difference between the actual value and the predicted value of the dependent variable. Unfortunately, in Cox regression there is no obvious analog this actual minus predicted. Realizing this, statisticians have looked at how residuals are used and then, based on those uses, developed quantities that meet those needs. They call these quantities *residuals* because they are used in place of residuals. However, you must remember that they are not equivalent to usual the residuals that you see in multiple regression, for example.

In the discussion that follows, the formulas will be simplified if we use the substitution

$$\theta_r = \exp\left(\sum_{i=1}^{p} x_{ir}\beta_i\right)$$

Cox-Snell Residuals

The Cox-Snell residuals were used to assess the goodness-of-fit of the Cox regression. The Cox-Snell residuals are defined as

$$r_t = H_{B0}(T_t)\theta_t$$

where there *b*'s are the estimated regression coefficients and $H_0(T_t)$ is Breslow's estimate of the cumulative baseline hazard function. This value is defined as follows

$$H_{B0}(T_t) = \sum_{T_i \le T_t}\left[\frac{m_i}{\sum_{j \in R_{T_i}} \theta_j}\right]$$

The Cox-Snell residuals were the first to be proposed in the literature. They have since been replaced by other types of residuals and are now only of historical interest. See, for example, the discussion of Marubini and Valsecchi (1996) who state that the use of these residuals on distributional grounds should be avoided.

Martingale Residuals

Martingale residuals can not be used to assess goodness-of-fit as are the usual residuals in multiple regression. The best model need not have the smallest sum of squared martingale residuals. Martingale residuals follow the unit exponential distribution. Some authors suggested analyzing these residuals to determine how close they are to the exponential distribution, hoping that a lack of exponetiality indicated a lack of fit. Unfortunately, just the opposite is the case since in a model with no useful covariates, these residuals are exactly exponential in distribution. Another diagnostic tool for in regular multiple regression is a plot of the residuals versus the fitted values. Here again, the martingale residuals cannot be used for this purpose since they are negatively correlated with the fitted values.

So of what use are martingale residuals? They have two main uses. First, they can be used to find outliers—individuals who are poorly fit by the model. Second, martingale residuals can be used to determine the functional form of each of the covariates in the model.

Finding Outliers

The martingale residuals are defined as

$$M_t = c_t - r_t$$

where c_t is one if there is a failure at time T_t and zero otherwise. The martingale residual measures the difference between whether an individual experiences the event of interest and the expected number of events based on the model. The maximum value of the residual is one and the minimum possible value is negative infinity. Thus, the residual is highly skewed. A large negative martingale residual indicates a high risk individual who still had a long survival time.

Finding the Function Form of Covariates

Martingale residuals can be used to determine the functional form of a covariate. To do this, you generate the Martingale residuals from a model without the covariates. Next, you plot these residuals against the value of the covariate. For large datasets, this may be a time consuming process. Therneau and Grambsch (2000) suggest that the martingale residuals from a model with no covariates be plotted against each of the covariates. These plots will reveal the appropriate functional form of the covariates in the model so long as the covariates are not highly correlated among themselves.

Deviance Residuals

Deviance residuals are used to search for outliers. The deviance residuals are defined as

$$DEV_t = sign(M_t)\sqrt{-2\left[M_t + c_t \ln(c_t - M_t)\right]}$$

or zero when M_t is zero. These residuals are plotted against the risk scores given by

$$\exp\left(\sum_{i=1}^{p} x_{it} b_i\right)$$

When there is slight to moderate censoring, large absolute values in these residuals point to potential outliers. When there is heavy censoring, there will be a large number of residuals near zero. However, large absolute values will still indicate outliers.

Schoenfeld's Residuals

A set of *p* Schoenfeld residuals is defined for each noncensored individual. The residual is missing when the individual is censored. The Schoenfeld residuals are defined as follows

$$r_{it} = c_t\left[x_{it} - \frac{\sum_{r\in R_t} x_{ir}\theta_r}{\sum_{r\in R_t}\theta_r}\right]$$
$$= c_t\left[x_{it} - \sum_{r\in R_t} x_{ir} w_r\right]$$

where

$$w_r = \frac{\sum_{r \in R_t} x_{ir}\theta_r}{\sum_{r \in R_t} \theta_r}$$

Thus this residual is the difference between the actual value of the covariate and a weighted average where the weights are determined from the risk scores.

These residuals are used to estimate the influence of an observation on each of the regression coefficients. Plots of these quantities against the row number or against the corresponding covariate values are used to study these residuals.

Scaled Schoenfeld's Residuals

Hosmer and Lemeshow (1999) and Therneau and Grambsch (2000) suggest that scaling the Schoenfeld residuals by an estimate of their variance gives quantities with greater diagnostic ability. Hosmer and Lemeshow (1999) use the covariance matrix of the regression coefficients to perform the scaling. The scaled Schoenfeld residuals are defined as follows

$$r_{kt}^* = m\sum_{i=1}^{p} V_{ik} r_{it}$$

where m is the total number of deaths in the dataset and V is the estimated covariance matrix of the regression coefficients.

These residuals are plotted against time to validate the proportional hazards assumption. If the proportional hazards assumption holds, the residuals will fall randomly around a horizontal line centered at zero. If the proportional hazards assumption does not hold, a trend will be apparent in the plot.

Data Structure

Survival data sets require up to three components for the survival time: the ending survival time, the beginning survival time during which the subject was not observed, and an indicator of whether the observation was censored or failed.

Based on these three components, various types of data may be analyzed. *Right censored* data are specified using only the ending time variable and the censor variable. *Left truncated* and *Interval* data are entered using all three variables.

The table below shows survival data ready for analysis. These data are from a lung cancer study reported in Kalbfleisch (1980), page 223. These data are in the LUNGCANCER database. The variables are

TIME	days of survival.
CENSOR	censor indicator.
STATUS	performance status.
MONTHS	months from diagnosis.
AGE	age in years.
THERAPY	prior therapy.

LUNGCANCER dataset (subset)

TIME	CENSOR	STATUS	MONTHS	AGE	THERAPY
72	1	60	7	69	0
411	1	70	5	64	10
228	1	60	3	38	0
126	1	60	9	63	10
118	1	70	11	65	10
10	1	20	5	49	0
82	1	40	10	69	10
110	1	80	29	68	0
314	1	50	18	43	0
100	0	70	6	70	0
42	1	60	4	81	0
8	1	40	58	63	10
144	1	30	4	63	0
25	0	80	9	52	10
11	1	70	11	48	10

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel lets you designate which variables are used in the analysis.

Time Variables

Time Variable

This variable contains the length of time that an individual was observed. This may represent a failure time or a censor time. Whether the subject actually died is specified by the Censor Variable. Since the values are elapsed times, they must be positive. Zeroes and negative values are treated as missing values.

During the maximum likelihood calculations, a risk set is defined for each individual. The risk set is defined to be those subjects who were being observed at this subject's failure and who lived as long or longer. It may take several rows of data to specify a subject's history.

This variable and the Entry Time Variable define a period during which the individual was at risk of failing. If the Entry Time Variable is not specified, its value is assumed to be zero.

Several types of data may be entered. These will be explained next.

- **Failure**

  This type of data occurs when a subject is followed from their entrance into the study until their death. The failure time is entered in this variable and the Censor Variable is set to the failed code, which is often a one.

  The Entry Time Variable is not necessary. If an Entry Time Variable is used, its value should be zero for this type of observation.

- **Interval Failure**

  This type of data occurs when a subject is known to have died during a certain interval. The subject may, or may not, have been observed during other intervals. If they were, they are treated as Interval Censored data. An individual may require several rows on the database to record their complete follow-up history.

  For example, suppose the condition of the subjects is only available at the end of each month. If a subject fails during the fifth month, two rows of data would be required. One row, representing the failure, would have a Time of 5.0 and an Entry Time of 4.0. The Censor variable would contain the failure code. A second row, representing the prior periods, would have a Time of 4.0 and an Entry Time of 0.0. The Censor variable would contain the censor code.

- **Censored**

  This type of data occurs when a subject has not failed up to the specified time. For example, suppose that a subject enters the study and does not die until after the study ends 12 months later. The subject's time (365 days) is entered here. The Censor variable contains the censor code.

- **Interval Censored**

  This type of data occurs when a subject is known not to have died during a certain interval. The subject may, or may not, have been observed during other intervals. An individual may require several rows on the database to record their complete follow-up history.

  For example, suppose the condition of the subjects is only available at the end of each month. If a subject fails during the fifth month, two rows of data would be required. One row, representing the failure, would have a Time of 5.0 and an Entry Time of 4.0. The Censor variable would contain the failure code. A second row, representing the prior periods, would have a Time of 4.0 and an Entry Time of 0.0. The Censor variable would contain the censor code.

Entry Time Variable

This optional variable contains the elapsed time before an individual entered the study. Usually, this value is zero. However, in cases such as *left truncation* and *interval censoring*, this value defines a time period before which the individual was not observed.

Negative entry times are treated as missing values. It is possible for the entry time to be zero.

Ties Method

The basic Cox regression model assumes that all failure times are unique. When ties exist among the failure times, one of two approximation methods is used to deal with the ties. When no ties are present, both of these methods result in the same estimates.

- **Breslow**

  This method was suggested first and is the default in many programs. However, the Efron method has been shown to be more accurate in most cases. The Breslow method is only used when you want to match the results of some other (older) Cox regression package.

- **Efron**

  This method has been shown to be more accurate, but requires slightly more time to calculate. This is the recommended method.

Censor Variable

Censor Variable

The values in this variable indicate whether the value of the Time Variable represents a censored time or a failure time. These values may be text or numeric. The interpretation of these codes is specified by the Failed and Censored options to the right of this option.

Only two values are used, the Failure code and the Censor code. The Unknown Type option specifies what is to be done with values that do not match either the Failure code or the Censor code.

Rows with missing values (blanks) in this variable are omitted from the estimation phase, but results are shown in any reports that output predicted values.

Failed

This value identifies those values of the Censor Variable that indicate that the Time Variable gives a failure time. The value may be a number or a letter.

We suggest the letter ‘F’ or the number ‘1’ when you are in doubt as to what to use.

A failed observation is one in which the time until the event of interest was measured exactly; for example, the subject died of the disease being studied. The exact failure time is known.

Left Censoring

When the exact failure time is not known, but instead only an upper bound on the failure time is known, the time value is said to have been *left censored*. In this case, the time value is treated as if it were the true failure time, not just an upper bound. So left censored observations should be coded as failed observations.

Censored

This value identifies those values of the Censor Variable that indicate that the individual recorded on this row was censored. That is, the actual failure time occurs sometime after the value of the Time Variable.

We suggest the letter ‘C’ or the number ‘0’ when you are in doubt as to what to use.

A censored observation is one in which the time until the event of interest is not known because the individual withdrew from the study, the study ended before the individual failed, or for some similar reason.

Note that it does not matter whether the censoring was Right or Interval. All you need to indicate here is that they were censored.

Unknown Censor

This option specifies what the program is to assume about observations whose censor value is not equal to either the Failed code or the Censored code. Note that observations with missing censor values are always treated as missing.

- **Censored**
  Observations with unknown censor values are assumed to have been censored.

- **Failed**
  Observations with unknown censor values are assumed to have failed.

- **Missing**

  Observations with unknown censor values are assumed to be missing and they are removed from the analysis.

Frequency Variable

Frequency Variable

This is an optional variable containing the frequency (observation count) for each row. Usually, you would leave this option blank and let each row receive the default frequency of one.

If your data have already been summarized, this option lets you specify how many actual rows each physical row represents.

Options

Centering of X's

The values of the independent variables may be centered to improve the stability of the algorithm. An value is 'centered' when its mean is subtracted from it.

Centering does not change the values of the regression coefficients, except that the algorithm might provide slightly different results because of better numerical stability.

Centering does affect the values of the row-wise statistics such as XB, Exp(XB), S0, H0, and so on because it changes the value of X in these expressions. When the data are centered, the deviation from the mean (X-Xbar) is substituted for X in these expressions.

The options are available:

- **None**

  The data are not centered.

- **All**

  All variables, both numeric and binary, are centered.

Alpha Level

Alpha is the significance level used in the hypothesis tests. One minus alpha is the confidence level of the confidence intervals. A value of 0.05 is most commonly used. This corresponds to a chance of error of 1 in 20. You should not be afraid to use other values since 0.05 became popular in pre-computer days when it was the only value available.

Typical values range from 0.001 to 0.20.

Numeric Independent Variables

X's: Numeric Independent Variables

Specify the numeric (continuous) independent variables. By numeric, we mean that the values are numeric and at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are better analyzed when you specify them as Categorical Independent Variables.

If you want to create powers and cross-products of these variables, specify an appropriate model in the 'Custom Model' field under the Model tab.

If you want to create hazard values for values of *X* not in your database, add the *X* values to the bottom of the database and leave their time and censoring blank. They will not be used during estimation, but various hazard and survival statistics will be generated for them and displayed in the Predicted Values report.

Categorical Independent Variables

X's: Categorical Independent Variable(s)

Specify categorical (nominal) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories.

The values in a categorical variable are not used directly in the regression analysis. Instead, a set of numeric variables is substituted for them. Suppose a categorical variable has *G* categories. ***NCSS*** automatically generates the *G*-1 indicator variables that are needed for the analysis. The type of indicator variable created is determined by the selection for the *Default Reference Value* and the *Default Contrast Type*. The type of indicator created can also be controlled by entering the reference value and contrast type directly according to the syntax below. See the Default Reference Value and Default Contrast Type sections below for a discussion of the reference value and contrast type options.

You can create the interactions among these variables automatically using the *Custom Model* field under the Model tab.

Syntax

The syntax for specifying a categorical variable is *VarName*(*RefValue*;*CType*) where *VarName* is the name of the variable, *RefValue* is the reference value, and *CType* is the type of numeric variables generated: B for binary, P for polynomial, R for contrast with the reference value, and S for a standard set of contrasts.For example, suppose a categorical variable, STATE, has four values: Texas, California, Florida, and New York. To process this variable, the values are arranged in sorted order: California, Florida, New York, and Texas. Next, the reference value is selected. If a reference value is not specified, the default value specified in the *Default Reference Value* window is used. Finally, the method of generating numeric variables is selected. If such a method is not specified, the contrast type selected in the *Default Contrast Type* window is used. Possible ways of specifying this variable are

STATE	**RefValue = Default, CType = Default**
STATE(New York)	**RefValue = New York, CType = Default**
STATE(California;R)	**RefValue = California, CType = Contrast with Reference**
STATE(Texas;S)	**RefValue = Texas, CType = Standard Set**

More than one category variable may be designated using a list. Examples of specifying three variables with various options are shown next.

STATE BLOODTYPE GENDER

STATE(California;R) BLOODTYPE(O) GENDER(F)

STATE(Texas;S) BLOODTYPE(O;R) GENDER(F;B)

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting**

  Use the first value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**

  Use the last value in alpha-numeric sorted order as the reference value.

The reference value may also be designated within parentheses after the name of the categorical independent variable, in which case the default reference value is ignored. For example, suppose that the categorical independent variable, STATE, has four values: 1, 3, 4, and 5.

1. If this option is set to 'First Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 1.
2. If this option is set to 'Last Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 5.
3. If the categorical independent variable is entered as 'STATE(4)', the choice for this setting would be ignored, and the reference value would be 4.

Default Contrast Type

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to something other than 'Binary'.

- **Binary (This is the default)**

  Categories are converted to numbers using a set of binary indicator variables by assigning a '1' to the active category and a '0' to all other values. For example, suppose a categorical variable has G categories. ***NCSS*** automatically generates the G-1 binary (indicator) variables that are used in the regression. These indicator variables are set to 1 for those rows in which the value of this variable is equal to a certain value. They are set to 0 otherwise. The G-1 occurs because the G^{th} indicator variable is redundant (when all G-1 indicators are 0, wIfe know that the G^{th} indicator variable would be a 1). The value that is skipped is called the Reference Value.

  If your model includes interactions, using the binary indicator type may cause strange results.

For the STATE variable, three binary variables would be generated. Suppose that the *Default Contrast Type* was 'Binary' and the statement used was 'STATE(Florida)'. The categories would be converted to numbers as follows:

STATE	B1	B2	B3
California	1	0	0
Florida	0	0	0
New York	0	1	0
Texas	0	0	1

- **Contrast with Reference**

  Categories are converted to numbers using a set of contrast variables by assigning a '1' to the active category, a '-1' to the reference value, and a '0' to all other values. A separate contrast is generated for each value other than the reference value.

  For the STATE variable, three numeric variables would be generated. Suppose the *Default Contrast Type* was 'Contrast with Reference', the *Default Reference Type* was 'Last Value after Sorting', and the variable was entered as 'STATE'. The categories would be converted to numbers as follows:

STATE	R1	R2	R3
California	1	0	0
Florida	0	1	0
New York	0	0	1
Texas	-1	-1	-1

- **Polynomial**

  If a variable has five or fewer categories, it can be converted to a set of polynomial contrast variables that account for the linear, quadratic, cubic, quartic, and quintic relationships. Note that these assignments are made after the values are sorted. Usually, the polynomial method is used on a variable for which the categories represent the actual values. That is, the values themselves are ordinal, not just category identifiers. Also, it is assumed that these values are equally spaced. Note that with this method, the reference value is ignored.

  For the STATE variable, linear, quadratic, and cubic variables are generated. Suppose that the *Default Contrast Type* was 'Polynomial' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	Linear	Quadratic	Cubic
California	-3	1	-1
Florida	-1	-1	3
New York	1	-1	-3
Texas	3	1	1

- **Standard Set**

  A variable can be converted to a set of contrast variables using a standard set of contrasts. This set is formed by comparing each value with those below it. Those above it are ignored. Note that these assignments are made after the values are sorted. The reference value is ignored.

For the STATE variable, three numeric variables are generated. Suppose that the *Default Contrast Type* was 'Standard Set' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	S1	S2	S3
California	-3	0	0
Florida	1	-2	0
New York	1	1	-1
Texas	1	1	1

Model Tab

These options control the regression model.

Subset Selection

Subset Selection

This option specifies the subset selection algorithm used to reduce the number of independent variables that used in the regression model. Note that since the solution algorithm is iterative, the selection process can be very time consuming. The Forward algorithm is much quicker than the Forward with Switching algorithm, but the Forward algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the generated individual binary variables. That is, either all binary variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **None**

  No subset selection is attempted. All specified independent variables are used in the regression equation.

- **(Hierarchical) Forward**

  With this algorithm, the term with the largest log likelihood is entered into the model. Next, the term that increases the log likelihood the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

  When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the log likelihood does not change significantly.

- **(Hierarchical) Forward with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest log likelihood is entered into the regression model. The term which increases the log likelihood the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, the likelihood function is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in the log likelihood. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Max Terms in Subset

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of the log likelihood.

Note that the intercept is counted in this number.

Estimation Options

These options control the number of iterations used while the algorithm is searching for the maximum likelihood solution.

Maximum Iterations

This option specifies the maximum number of iterations used while finding a solution. If this number is reached, the procedure is terminated prematurely. This is used to prevent an infinite loop and to reduce the running time of lengthy variable selection runs.

Usually, no more the 20 iterations are needed. In fact, most runs converge in about 7 or 8 iterations.

During a variable selection run, it may be advisable reset this value to 4 or 5 to speed up the variable selection. Usually, the last few iterations make little difference in the estimated values of the regression coefficients.

Convergence Zero

This option specifies the convergence target for the maximum likelihood estimation procedure. The algorithm finds the maximum relative change of the regression coefficients. If this amount is less than the value set here, the maximum likelihood procedure is terminated.

For large datasets, you might want to increase this value to about 0.0001 so that fewer iterations are used, thus decreasing the running time of the procedure.

Model Specification

Which Model Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a straight-forward regression model, select *Up to 1-Way*.

The options are

- **Full Model**

  The complete, saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables).

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

  Note that the discussion of the Custom Model option discusses the interpretation of this model.

- **Up to 1-Way**

  This option generates a model in which each variable is represented by a single model term. No cross-products or interaction terms are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

  This is the option to select when you want to analyze the independent variables specified without adding any other terms.

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C

- **Up to 2-Way**

  This option specifies that all main effects and two-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C

- **Up to 3-Way**

  All main effects, two-way interactions, and three-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

- **Up to 4-Way**

  All main effects, two-way interactions, three-way interactions, and four-way interactions are included in the model. For example, if you have four independent variables A, B, C, and D, this would generate the model:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D

- **Custom Model**

  The model specified in the *Custom Model* box is used.

Write Model in Custom Model Field

When this option is checked, no data analysis is performed when the procedure is run. Instead, a copy of the full model is stored in the Custom Model box. You can then edit the model as desired. This option is useful when you want to be selective about which terms to keep and you have several variables.

Note that the program will not do any calculations while this option is checked.

Model Specification – Custom Model

Max Term Order

This option specifies that maximum number of variables that can occur in an interaction term in a custom model. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would be excluded from the model.

This option is particularly useful when used with the bar notation of a custom model to allow a simple way to remove unwanted high-order interactions.

Custom Model

This options specifies a custom model. It is only used when the *Which Model Terms* option is set to *Custom Model*. A custom model specifies the terms (single variables and interactions) that are to be kept in the model.

Interactions

An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of indicator variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the indicator variables generated from the categorical variable.

Syntax

A model is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A + B + C + A*B + A*C + B*C + A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B + A*C.

Some examples will help to indicate how the model syntax works:

A|B = A + B + A*B

A|B A*A B*B = A + B + A*B + A*A + B*B

Note that you should only repeat numeric variables. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A + B + A*A + A*B + B*B

A|B|C = A + B + C + A*B + A*C + B*C + A*B*C

(A + B)*(C + D) = A*C + A*D + B*C + B*D

(A + B)|C = (A + B) + C + (A + B)*C = A + B + C + A*C + B*C

Reports Tab

The following options control which reports are displayed.

Select Reports – Summaries

Run Summary

Indicate whether to display this summary report.

Select Reports – Subset Selection

Subset Selection - Summary and Subset Selection - Detail

Indicate whether to display these subset selection reports.

Select Reports – Estimation

Regression Coefficients ... C.L. of Regression Coefficients

Indicate whether to display these estimation reports.

Select Reports – Goodness-of-Fit

Analysis of Deviance ... Baseline Hazard and Survival

Indicate whether to display these model goodness-of-fit reports.

Select Reports – Row-by-Row Lists

Residuals ... Predicted Values

Indicate whether to display these list reports. Note that since these reports provide results for each row, they may be too long for normal use when requested on large databases.

Order of Row Reports

This option specifies the order of the observations displayed on reports that display a separate value for each row. The rows can be displayed in the original order of the database or sorted by the time value, from lowest to highest.

Select Plots

Null Martingale Resid vs X Plot ... Deviance Resid vs Time Plot

Indicate whether to display these plots.

Format Tab

These options control format of the reports.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Skip Line After

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Report Options – Decimal Places

Time ... Z or Chi2 Decimals

These options specify the number of decimal places shown on the reports for the indicated values.

MResid vs X Plots to Resid vs Time Plots Tabs

These options control the attributes of the various plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful..

Data Storage Options – Select Items to Store

Expanded X Values ... Covariance Matrix

Indicate whether to store these row-by-row values, beginning at the variable indicated by the *Store First Variable In* option. Note that several of these values include a different value for each covariate and so they require several columns when they are stored.

Expanded X Values

This option refers to the experimental design matrix. They include all binary and interaction variables generated.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Cox Regression Analysis

This section presents an example of how to run a Cox regression analysis. The data used are found in the LUNGCANCER database. These data are an excerpt from a lung cancer study reported in Kalbfleisch (1980). The variables used in the analysis are

TIME	Days of survival
CENSOR	Censor indicator
STATUS	Karnofsky rating performance status
MONTHS	Months from diagnosis
AGE	Age in years
THERAPY	Prior therapy: 0 no, 10 yes

The purpose of this analysis is study the relationship between length of patient survival and the covariates. You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Cox Regression window.

1 Open the LUNGCANCER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LungCancer.s0**.
- Click **Open**.

2 Open the Cox Regression window.

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Cox Regression**. The Cox Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will load the default template.

3 Specify the variables.

- On the Cox Regression window, select the **Variables tab**.
- Enter **Time** in the **Time Variable** box.
- Set the **Ties Method** to **Efron**.
- Enter **Censor** in the **Censor Variable** box.
- Enter **Status-Therapy** in the **X's: Numeric Independent Variables** box.

4 Specify the reports.

- On the Cox Regression window, select the **Reports tab**.
- Check all of the reports. Although under normal circumstances you would not need all of the reports, we will view them all here so they can be annotated.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Section

Run Summary Section

Parameter	Value	Parameter	Value
Rows Read	15	Time Variable	Time
Rows Filtered Out	0	Censor Variable	Censor
Rows Missing X's	0	Frequency Variable	None
Rows Processed	15	Subset Method	None
Rows Prediction Only	0	Ind. Var's Available	4
Rows Failed	13	No. of X's in Model	4
Rows Censored	2	Iterations	7
Sum of Frequencies	15	Maximum Iterations	20
Sum Censored Freqs	2	Convergence Criterion	1E-09
Sum Failed Freqs	13	Achieved Convergence	1.473012E-15
Final Log Likelihood	-20.1143	Completion Message	Normal completion

This report summarizes the characteristics of the dataset and provides useful information about the reports to follow. It should be studied to make sure that the data were read in properly and that the estimation algorithm terminated normally. We will only discuss those parameters that need special explanation.

Rows Read

This is the number of rows processed during the run. Check this count to make certain it agrees with what you anticipated.

Iterations

This is the number of iterations used by the maximum likelihood procedure. This value should be compared against the value of the Maximum Iterations option to see if the iterative procedure terminated early.

Achieved Convergence

This is the maximum of the relative changes in the regression coefficients on the last iteration. If this value is less than the Convergence Criterion, the procedure converged normally. Otherwise, the specified convergence precision was not achieved.

Final Log Likelihood

This is the log likelihood of the model.

Regression Coefficients Section

Regression Coefficients Section

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	Prob Level	Pseudo R2
B1 Age	0.039805	0.035232	1.0406	60.3333	1.1298	0.2586	0.1242
B2 Months	0.064557	0.033056	1.0667	12.6000	1.9530	0.0508	0.2977
B3 Status	-0.032415	0.020324	0.9681	57.3333	-1.5949	0.1107	0.2204
B4 Therapy	0.013967	0.068384	1.0141	4.6667	0.2042	0.8382	0.0046

Estimated Cox Regression Model

Exp(3.98048128120681E-02 + 6.45571159984993E-02*Months -3.24152392634531E-02*Status
+ 1.39668973406698E-02*Therapy)

This report displays the results of the proportional hazards estimation. Following are the detailed definitions:

Independent Variable

This is the variable from the model that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the binary variable is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the Skip Line After option of the Format tab.

Regression Coefficient (B)

This is the estimate of the regression coefficient, β_i. Remember that the basic regression equation is

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^{p} x_i \beta_i$$

Thus the quantity β_i is the amount that the log of the hazard rate changes when x_i is increased by one unit. Note that a positive coefficient implies that as the value of the covariate is increased, the hazard increases and the prognosis gets worse. A negative coefficient indicates that as the variable is increased, the hazard decreases and the prognosis gets better.

Standard Error

This is s_{b_j}, the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is provided by the square root of the corresponding diagonal element of the covariance matrix, $V(\hat{\beta}) = I^{-1}$. It is also used as the denominator of the Wald test.

Risk Ratio Exp(B)

This the value of e^{β_i}. This value is often called the *risk ratio* since it is the ratio of two hazards whose only difference is that x_i is increased by one unit. That is,

$$\frac{h(T \mid x_i = a+1)}{h(T \mid x_i = a)} = e^{\beta_i}$$

In this example, if Months is increased by one, the hazard rate is increased by 6.67%. If you want to calculate the affect of increasing Months by three, the hazard rate is increased by $1.0667^3 = 1.2137$, or 21.37%. Note that is not equal to 3.0 times 6.67.

Mean

This is the average of this independent variable. The means are especially important in interpreting the baseline hazard rates. Unless you have opted otherwise, the independent variables have been centered by subtracting these mean values. Hence, the baseline hazard rate occurs when each independent variable is equal to its mean.

Wald Z-Value

This is the *z* value of the Wald test used for testing the hypothesis that $\beta_i = 0$ against the alternative $\beta_i \neq 0$. The Wald test is calculated using the formula

$$z_i = \frac{b_{ij}}{s_{b_i}}$$

The distribution of the Wald statistic is closely approximated by the normal distribution in large samples. However, in small samples, the normal approximation may be poor. For small samples, likelihood ratio tests perform better and are preferred.

Prob Level

This is the two-sided probability level. This is the probability of obtaining a z-value larger in absolute value than the one obtained. If this probability is less than the specified significance level (say 0.05), the regression coefficient is significantly different from zero.

Pseudo R2

An index value, similar to R-Squared in regression, representing the relative influence of this variable. If *C = z^2*, *n* = sample size, and *p* = number of variables, then *R2 = C/(n-p+C)*.

Estimated Cox Model

This section gives the Cox regression model in a regular text format that can be used as a transformation formula. The regression coefficients are displayed in double precision because a single-precision formula does not include the accuracy necessary to calculate the hazard rates.

Note that transformation must be less than 255 characters. Since these formulas are often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file, including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

Confidence Limits Section

Confidence Limits Section

Independent Variable	Regression Coefficient (B)	Lower 95.0% Confidence Limit of B	Upper 95.0% Confidence Limit of B	Risk Ratio Exp(B)	Lower 95.0% C.L. of Exp(B)	Upper 95.0% C.L. of Exp(B)
B1 Age	0.039805	-0.029249	0.108858	1.0406	0.9712	1.1150
B2 Months	0.064555	-0.000231	0.129341	1.0667	0.9998	1.1381
B3 Status	-0.032415	-0.072249	0.007420	0.9681	0.9303	1.0074
B4 Therapy	0.013965	-0.120064	0.147993	1.0141	0.8869	1.1595

This report provides the confidence intervals for the regression coefficients and the risk ratios. The confidence coefficient, in this example 95%, was specified on the Format tab.

Independent Variable

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the Skip Line After option of the Format tab.

Regression Coefficient (B or Beta)

This is the estimate of the regression coefficient, β_i. Remember that the basic regression equation is

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^{p} x_i \beta_i$$

Thus the quantity β_i is the amount that the log of the hazard rate changes when x_i is increased by one unit. Note that a positive coefficient implies that as the value of the covariate is increased, the hazard increases and the prognosis gets worse. A negative coefficient indicates that as the variable is increased, the hazard decreases and the prognosis gets better.

Confidence Limits of B

A 95% confidence interval for β_i is given by an upper and lower limit. These limits are based on the Wald statistic using the formula

$$b_i \pm z_{1-\alpha/2} s_{b_i}$$

Since they are based on the Wald test, they are only valid for large samples.

Risk Ratio Exp(B)

This the value of e^{β_i}. This value is often called the *risk ratio* since it is the ratio of two hazards whose only difference is that x_i is increased by one unit. That is,

$$\frac{h(T \mid x_i = a+1)}{h(T \mid x_i = a)} = e^{\beta_i}$$

In this example, if Months is increased by one, the hazard rate is increased by 6.67%. If you want to calculate the affect of increasing Months by three, the hazard rate is increased by $1.0667^3 = 1.2137$, or 21.37%. Note that is not equal to 3.0 times 6.67.

Confidence Limits of Exp(B)

A 95% confidence interval for e^{β_i} is given by an upper and lower limit. These limits are based on the Wald statistic using the formula

$$\exp\left(b_i \pm z_{1-\alpha/2} s_{b_i}\right)$$

Since they are based on the Wald test, they are only valid for large samples.

Analysis of Deviance Section

Term(s) Omitted	DF	-2 Log Likelihood	Increase From Model Deviance (Chi Square)	Prob Level
All Terms	4	46.6698	6.4413	0.1685
AGE	1	41.5943	1.3657	0.2426
MONTHS	1	44.3928	4.1642	0.0413
STATUS	1	42.7787	2.5501	0.1103
THERAPY	1	40.2704	0.0419	0.8379
None(Model)	4	40.2286		

This report is the Cox regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms in the regression are statistically significant after adjusting for all other terms in the model.

This report is not produced during a subset selection run.

Note that this report requires that a separate regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.

The “All” line refers to a no-covariates model. The “None(Model)” refers to the complete model with no terms removed.

Note that it is usually not advisable to include an interaction term in a model when one of the associated main effects is missing—which is what happens here. However, in this case, we believe this to be a useful test.

Note also that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the chi-square test displayed on this line. DF is equal to the number of individual independent variables in the term.

Log Likelihood

This is the log likelihood achieved by the model being described on this line of the report.

R-Squared of Remaining Terms

This is the difference between the deviance for the model described on this line and the deviance of the complete model. This value follows the chi-square distribution in medium to large samples. This value can be thought of as the analog of the residual sum of squares in multiple regression. Thus, you can think of this value as the increase in the residual sum of squares that occurs when this term is removed from the model.

Another way to interpret this test is as a *redundancy test* because it tests whether this term is redundant after considering all of the other terms in the model.

Prob Level

This is the significance level of the chi-square test. This is the probability that a chi-square value with degrees of freedom DF is equal to this value or greater. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

Log Likelihood & R-Squared Section

Term(s) Omitted	DF	Log Likelihood	R-Squared Of Remaining Term(s)	Reduction From Model R-Squared
All Terms	4	-23.3349	0.0000	0.3491
AGE	1	-20.7971	0.2871	0.0620
MONTHS	1	-22.1964	0.1408	0.2083
STATUS	1	-21.3893	0.2285	0.1206
THERAPY	1	-20.1352	0.3473	0.0018
None(Model)	4	-20.1143	0.3491	0.0000

This report provides the log likelihoods and R-squared values of various models. This report is not produced during a subset selection run.

Note that this report requires that a separate Cox regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the term that is omitted from the model. The "All" line refers to no-covariates model. The "None(Model)" refers to the complete model with no terms removed. The "None(Model)" refers to the complete model with no terms removed.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the term displayed on this line.

Log Likelihood

This is the log likelihood of the model displayed on this line. Note that this is the log likelihood of the logistic regression without the term listed.

R-Squared of Remaining Term(s)

This is the *R*-squared of the model displayed on this line. Note that the model does not include the term listed at the beginning of the line. This *R*-squared is analogous to the *R*-squared in multiple regression, but it is not the same.

Hosmer and Lemeshow (1999) indicate that at the time of the writing of their book, there is no single, easy to interpret measure in Cox regression that is analogous to R^2 in multiple regression. They indicate that if such a measure "must be calculated" they would use

$$R_p^2 = 1 - \exp\left[\frac{2}{n}\left(L_0 - L_p\right)\right]$$

where L_0 is the log likelihood of the model with no covariates, *n* is the number of observations (censored or not), and L_p is the log likelihood of the model that includes the covariates.

Reduction From Model R-Squared

This is amount that *R*-squared is reduced when the term is omitted from the regression model. This reduction is calculated from the *R*-squared achieved by the full model.

This quantity is used to determine if removing a term causes a large reduction in *R*-squared. If it does not, then the term can be safely removed from the model.

Baseline Cumulative Hazard & Survival Section

Time	Centered Baseline Cumulative Survival	Centered Baseline Cumulative Hazard	Alpha	Centered Baseline Hazard Rate
8	0.9654	0.0352	0.9654	0.0346
10	0.8912	0.1152	0.9232	0.0768
11	0.8183	0.2006	0.9181	0.0819
42	0.7449	0.2945	0.9103	0.0897
72	0.6717	0.3980	0.9017	0.0983
82	0.5934	0.5220	0.8834	0.1166
110	0.4942	0.7048	0.8329	0.1671
118	0.3904	0.9407	0.7898	0.2102
126	0.2911	1.2341	0.7457	0.2543
144	0.1843	1.6915	0.6330	0.3670
228	0.0922	2.3841	0.5003	0.4997
314	0.0288	3.5461	0.3128	0.6872
411	0.0288	3.5461	0.0000	1.0000

This report displays various estimated survival and hazard values. These are centered if the Centered X's option is selected.

Baseline Cumulative Survival

This estimates the cumulative survival probability of an individual with all covariates equal to their means or to zero depending on whether the data are centered or not. It is the value of $S_0(T)$ which is estimated using the formula

$$S_0(T) = \prod_{T_t \le T} \alpha_t$$

Baseline Cumulative Hazard

This estimates the cumulative baseline hazard of an individual with a set of covariates all equal to zero. It is the value of $H_0(T)$ which is calculated using the formula

$$H_0(T) = -\ln(S_0(T))$$

Alpha

This is the value of the conditional baseline survival probabilities at the times listed. These values are used to calculate $S_0(T)$.

Baseline Hazard Rate

This is the estimate of the baseline hazard rates $h_0(T_t)$ which are calculated as follows

$$h_0(T_t) = 1 - \alpha_t$$

Residual Section

Row	Time	Cox-Snell Residual		Martingale Residual		Deviance Residual	
12	8	1.3861	||||||.....	-0.3861	||.....	-0.3453	||.....
6	10	0.1411	|.....	0.8589	||||||.....	1.4828	||||||||||...
15	11	0.0791	|.....	0.9209	||||||.....	1.7978	||||||||||||||
14+	25	0.0590	|.....	-0.0590	|.....	-0.3434	||.....
11	42	0.3307	|.....	0.6693	||||.....	0.9351	|||||||.....
1	72	0.3364	|.....	0.6636	||||.....	0.9229	|||||||.....
7	82	1.1774	|||||.....	-0.1774	|.....	-0.1679	|.....
10+	100	0.3112	|.....	-0.3112	|.....	-0.7890	||||||.....
8	110	1.2387	||||||.....	-0.2387	|.....	-0.2220	|.....
5	118	0.7300	|||.....	0.2700	|.....	0.2991	|.....
4	126	1.0748	|||||.....	-0.0748	|.....	-0.0730	|.....
13	144	2.4532	||||||||||...	-1.4532	|||||||||.....	-1.0543	|||||||.....
3	228	0.4532	||.....	0.5468	|||.....	0.6995	|||||.....
9	314	2.9953	||||||||||||||	-1.9953	||||||||||||||	-1.3403	||||||||||....
2	411	1.7951	||||||||.....	-0.7951	|||||.....	-0.6481	|||||.....

The various residuals were discussed in detail earlier in this chapter. Only a brief definition will be given were.

Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

Time

This is the value of the elapsed time.

Cox-Snell Residuals

Cox-Snell residuals were created to assess the goodness-of-fit of the Cox regression. They have since been replaced by other types of residuals and are now only of historical interest. See, for example, the discussion of Marubini and Valsecchi (1996) who state that the use of these residuals on distributional grounds should be avoided.

Martingale Residuals

The martingale residuals are defined as

$$M_t = c_t - r_t$$

where c_t is one if there is a failure at time T_t and zero otherwise. The martingale residual measures the difference between whether an individual experiences the event of interest and the expected number of events based on the model. The maximum value of the residual is one and the minimum possible value is negative infinity. Thus, the residual is highly skewed. A large negative martingale residual indicates a high risk individual who still had a long survival time.

Martingale residuals can not be used to assess goodness-of-fit as are the usual residuals in multiple regression. They have two main uses. First, they can be used to find outliers—individuals who are poorly fit by the model. Second, martingale residuals can be used to determine the functional form of each of the covariates in the model.

Martingale residuals can be used to determine the functional form of a covariate. To do this, you generate the Martingale residuals from a model without the covariate. Next, you plot these residuals against the value of the covariate.

Deviance Residuals

Deviance residuals are used to search for outliers. The deviance residuals are defined as

$$DEV_t = sign(M_t)\sqrt{-2[M_t + c_t \ln(c_t - M_t)]}$$

or zero when M_t is zero. These residuals are plotted against the risk scores given by

$$\exp\left(\sum_{i=1}^{p} x_{it} b_i\right)$$

When there is slight to moderate censoring, large absolute values in these residuals point to potential outliers. When there is heavy censoring, there will be a large number of residuals near zero. However, large absolute values will still indicate outliers.

Martingale Residuals Section

Row	Time	Null Martingale Residual		Martingale Residual	
12	8	0.9310	|||||||.....	-0.3862	||.....
6	10	0.8569	|||||||.....	0.8589	|||||.....
15	11	0.7769	|||||||.....	0.9209	|||||.....
14+	25	-0.2231	|.....	-0.0590	|.....
11	42	0.6815	|||||.....	0.6693	||||.....
1	72	0.5762	|||||.....	0.6636	||||.....
7	82	0.4584	|||.....	-0.1774	|.....
10+	100	-0.5416	||||.....	-0.3112	|.....
8	110	0.3043	||.....	-0.2387	|.....
5	118	0.1219	|.....	0.2700	|.....
4	126	-0.1012	|.....	-0.0748	|.....
13	144	-0.3889	|||.....	-1.4532	|||||||||.....
3	228	-0.7944	|||||||.....	0.5468	|||.....
9	314	-1.4875	||||||||||||||	-1.9953	||||||||||||||
2	411	-1.4875	||||||||||||||	-0.7951	|||||.....

The various residuals were discussed in detail earlier in this chapter. Only a brief definition will be given were.

Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

Time

This is the value of the elapsed time.

Null Martingale Residuals

These are the null martingale residuals. They are computed from a null (no covariate) model. Therneau and Grambsch (2000) suggest that the null-model martingale residuals can show the ideal functional form of the covariates so long as the covariates are not highly correlated among themselves. To find the appropriate functional form, each covariate is plotted against these residuals.

Martingale Residuals

The martingale residuals are repeated here. They were defined in the Residuals Section.

Schoenfeld Residuals Section

Schoenfeld Residuals Section

Row	Time	Resid Age		Resid Months		Resid Status	
12	8	-0.1121	|.....	11.8151	||||||||||...	-3.4330	|.....
6	10	-14.4483	||||||||||...	-5.7471	|||||.....	-33.7298	||||||||||||
15	11	-16.9356	||||||||||||	-0.3387	|.....	12.7982	|||||.....
11	42	15.1298	|||||||||||..	-7.4119	|||||||.....	3.8457	|.....
1	72	4.8130	||||.....	-5.2365	|||||.....	4.2736	|.....
7	82	5.2529	||||.....	-2.7151	||.....	-15.3358	||||||.....
8	110	6.6884	|||||.....	14.7013	|||||||||||||	20.6226	||||||||.....
5	118	6.2230	|||||.....	2.2723	||.....	18.4375	||||||||.....
4	126	5.4733	||||.....	0.7289	|.....	12.1419	|||||.....
13	144	7.0668	||||||.....	-4.0589	||||.....	-14.3231	||||||.....
3	228	-11.2819	||||||||.....	-8.8791	||||||||.....	2.1963	|.....
9	314	-7.8694	||||||.....	4.8715	||||.....	-7.4947	|||.....
2	411	0.0000	|.....	0.0000	|.....	0.0000	|.....

Schoenfeld Residuals Section

Row	Time	Resid Therapy	
12	8	1.5297	|||.....
6	10	-3.8822	||||||||.....
15	11	5.7182	|||||||||||..
11	42	-3.9309	||||||||.....
1	72	-4.3682	|||||||||.....
7	82	5.2326	||||||||||...
8	110	-3.3664	|||||||.....
5	118	5.3579	||||||||||...
4	126	6.4343	|||||||||||||
13	144	-1.6923	|||.....
3	228	-3.2851	|||||||.....
9	314	-3.7473	|||||||.....

This report displays the Schoenfeld residuals for each noncensored individual. Note that most authors suggest using the scaled Schoenfeld residuals rather than these residuals. Since these residuals were discussed earlier in this chapter, only a brief definition will be given were.

Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

Time

This is the value of the elapsed time.

Schoenfeld Residuals

The Schoenfeld residuals are defined as follows

$$r_{it} = c_t \left[x_{it} - \sum_{r \in R_t} x_{ir} w_r \right]$$

where

$$w_r = \frac{\sum_{r \in R_t} x_{ir} \theta_r}{\sum_{r \in R_t} \theta_r}$$

Thus, this residual is the difference between the actual value of the covariate and a weighted average where the weights are determined from the risk scores. These residuals are used to estimate the influence of an observation on each of the regression coefficients. Plots of these quantities against the row number or against the corresponding covariate values are used to study these residuals.

Scaled Schoenfeld Residuals Section

Scaled Schoenfeld Residuals Section

Row	Time	Resid Age		Resid Months		Resid Status	
12	8	0.0276	|.....	0.1828	|||||||||||||	-0.0569	|||||.....
6	10	-0.1660	|||||||.....	-0.0528	||||.....	-0.1280	|||||||||||||
15	11	-0.3476	|||||||||||||	-0.0731	|||||.....	0.0701	|||||||.....
11	42	0.2516	|||||||||.....	-0.0812	|||||.....	0.0355	||||.....
1	72	0.0959	||||.....	-0.0873	||||||.....	0.0487	|||||.....
7	82	0.0482	||.....	0.0410	|||.....	-0.1048	||||||||||...
8	110	0.1586	|||||.....	0.1616	|||||||||||..	0.0753	|||||||.....
5	118	0.0300	|.....	0.0237	|.....	0.0611	||||||.....
4	126	0.0116	|.....	0.0206	|.....	0.0280	|||.....
13	144	0.1376	|||||.....	-0.0020	|.....	-0.0690	|||||||.....
3	228	-0.1832	|||||||.....	-0.1822	||||||||||||.	0.0663	||||||.....
9	314	-0.0642	||.....	0.0489	||||.....	-0.0262	|||.....
2	411	0.0000	|.....	0.0000	|.....	0.0000	|.....

Scaled Schoenfeld Residuals Section

Row	Time	Resid Therapy	
12	8	0.1550	|||||.....
6	10	0.0223	|.....
15	11	0.4537	|||||||||||||
11	42	-0.4288	||||||||||||.
1	72	-0.3501	|||||||||....
7	82	0.3223	|||||||||.....
8	110	-0.2982	||||||||.....
5	118	0.1971	|||||.....
4	126	0.2903	||||||||.....
13	144	-0.1256	||||.....
3	228	-0.1361	||||.....
9	314	-0.1018	|||.....
2	411	0.0000	|.....

This report displays the scaled Schoenfeld residuals for each noncensored individual. These residuals are often used to find influential observations. Since these residuals were discussed earlier in this chapter, only a brief definition will be given were.

Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

Time

This is the value of the elapsed time.

Scaled Schoenfeld Residuals

The scaled Schoenfeld residuals are defined as follows

$$r_{kt}^* = m\sum_{i=1}^{p} V_{ik} r_{it}$$

where *m* is the total number of deaths in the dataset and *V* is the estimated covariance matrix of the regression coefficients. Hosmer and Lemeshow (1999) and Therneau and Grambsch (2000) suggest that scaling the Schoenfeld residuals by an estimate of their variance gives quantities with greater diagnostic ability. Hosmer and Lemeshow (1999) use the covariance matrix of the regression coefficients to perform the scaling.

These residuals are plotted against time to validate the proportional hazards assumption. If the proportional hazards assumption holds, the residuals will fall randomly around a horizontal line centered at zero. If the proportional hazards assumption does not hold, a trend will be apparent in the plot.

Predicted Values Section

Row	Time	Cumulative Baseline Hazard	Linear Predictor XB	Relative Risk Exp(XB)	Cumalitive Hazard H(T|X)	Cumulative Survival S(T|X)
12	8	0.0352	3.6734	39.3805	1.3861	0.2500
6	10	0.1152	0.2032	1.2254	0.1411	0.8684
15	11	0.2006	-0.9303	0.3944	0.0791	0.9239
14+	25	0.2006	-1.2244	0.2939	0.0590	0.9427
11	42	0.2945	0.1158	1.1228	0.3307	0.7184
1	72	0.3980	-0.1681	0.8452	0.3364	0.7143
7	82	0.5220	0.8135	2.2557	1.1774	0.3081
10+	100	0.5220	-0.5170	0.5963	0.3112	0.7325
8	110	0.7048	0.5640	1.7576	1.2387	0.2898
5	118	0.9407	-0.2536	0.7760	0.7300	0.4819
4	126	1.2341	-0.1382	0.8709	1.0748	0.3414
13	144	1.6915	0.3718	1.4503	2.4532	0.0860
3	228	2.3840	-1.6603	0.1901	0.4532	0.6356
9	314	3.5461	-0.1688	0.8447	2.9953	0.0500
2	411	3.5461	-0.6808	0.5062	1.7951	0.1661

This report displays various values estimated by the model. These are centered if the Centered X's option is selected.

Row

This is the row from the database that is displayed on this line. Rows with a plus sign were censored.

Time

This is the value of the elapsed time.

Baseline Cumulative Hazard

This estimates the cumulative baseline hazard of this individual. The baseline hazard occurs when all covariates are equal to zero (or to their means if centering is used). It is the value of $H_0(T)$ which is calculated using the formula

$$H_0(T) = -\ln(S_0(T))$$

Linear Predictor

This is the value of the linear portion of the Cox regression model. It is the logarithm of the ratio of the hazard rate to the baseline hazard rate. That is, it is the logarithm of the hazard ratio (or relative risk). The formula for the linear predictor is

$$\ln\left[\frac{h(T)}{h_0(T)}\right] = \sum_{i=1}^{p} x_i \beta_i$$

This value is occasionally suggested for use in plotting.

Relative Risk Exp(XB)

This is the ratio between the actual hazard rate and the baseline hazard rate, sometimes called the *risk ratio* or the *relative risk*. The formula for this quantity is

$$\begin{aligned}\frac{h(T)}{h_0(T)} &= \exp\left(\sum_{i=1}^{p} x_i\beta_i\right) \\ &= e^{x_1\beta_1}e^{x_2\beta_2}\cdots e^{x_p\beta_p}\end{aligned}$$

Cumulative Hazard H(T|X)

Under the proportional hazards regression model, the cumulative hazard is the sum of the individual hazard rates from time zero to time *T*.

$$\begin{aligned}H(T,X) &= \int_0^T h(u,X)du \\ &= \int_0^T h_0(u)e^{\sum_{i=1}^{p} x_i\beta_i}du \\ &= e^{\sum_{i=1}^{p} x_i\beta_i}\int_0^T h_0(u)du \\ &= H_0(T)e^{\sum_{i=1}^{p} x_i\beta_i}\end{aligned}$$

Note that the time survival time *T* is present in $H_0(T)$, but not in $e^{\sum_{i=1}^{p} x_i\beta_i}$. Hence, the cumulative hazard up to time *T* is represented in this model by a baseline cumulative hazard $H_0(T)$ which is adjusted for the covariates by multiplying by the factor $e^{\sum_{i=1}^{p} x_i\beta_i}$.

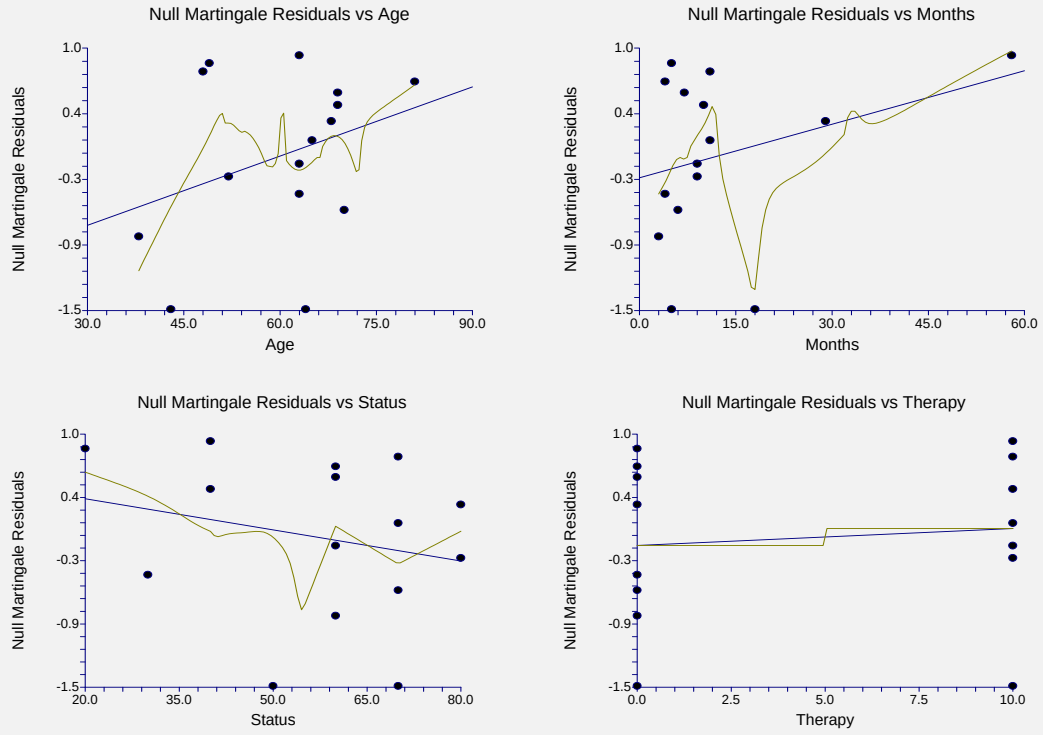
Cumulative Survival S(T|X)

Under the proportional hazards regression model, the cumulative survival is the probability that an individual survives past *T*. The formula for the cumulative survival is

$$\begin{aligned}S(T,X) &= \exp(-H(T,X)) \\ &= \exp\left(-H_0(T)e^{\sum_{i=1}^{p} x_i\beta_i}\right) \\ &= \left[e^{-H_0(T)}\right]^{e^{\sum_{i=1}^{p} x_i\beta_i}} \\ &= S_0(T)^{e^{\sum_{i=1}^{p} x_i\beta_i}}\end{aligned}$$

Note that the time survival time *T* is present in $S_0(T)$, but not in $e^{\sum_{i=1}^{p} x_i\beta_i}$.

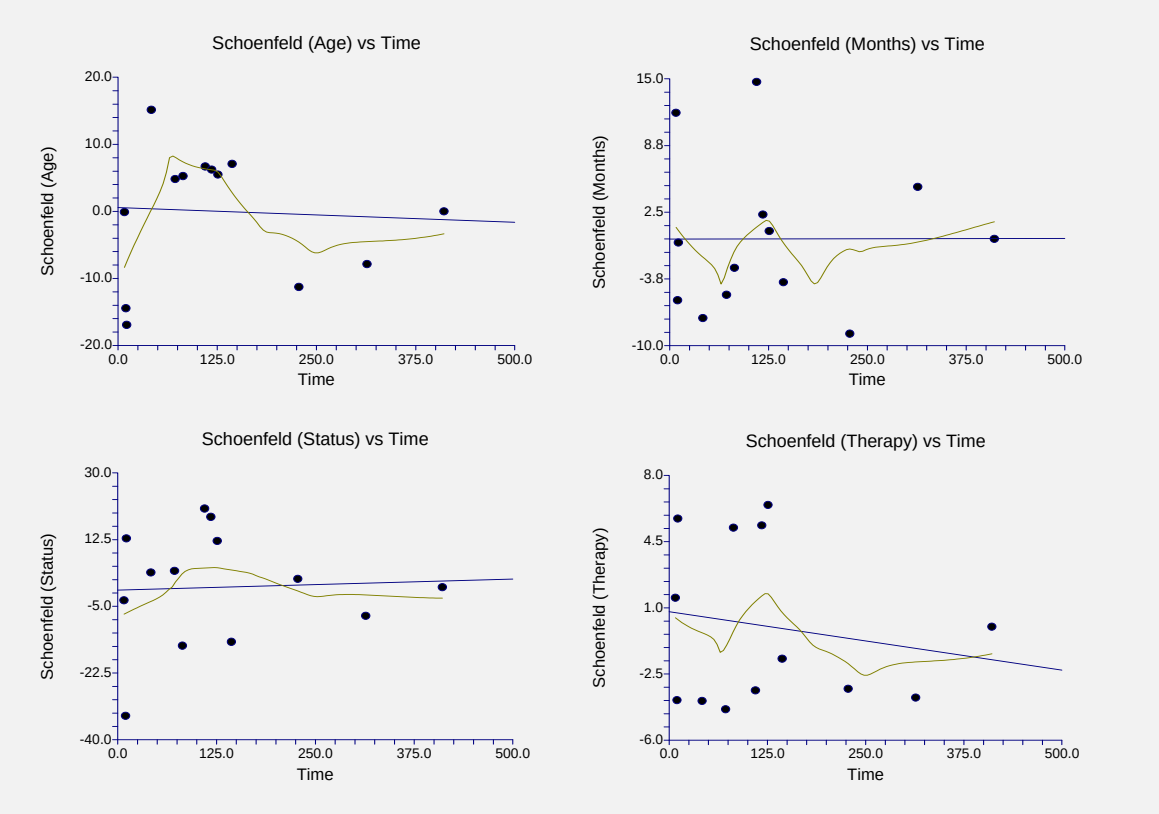
Plots of Null Martingale Residuals Versus Each of the Covariates



Each of the covariates are plotted against the null martingale residuals. If the covariates are not highly correlated, these plot will show the appropriate functional form of each covariate. A lowess curve and a regular least squares line are added to the plot to aid the eye. Ideally, the lowess curve will track along the least squares line. Be careful not to over interpret the ends of the lowess curves which are based on only a few individuals.

When curvature is present, you have to decide how the model should be modified to deal with it. You might need to add the square or the logarithm of the covariate to the model.

Plots of Schoenfeld Residuals Versus Time

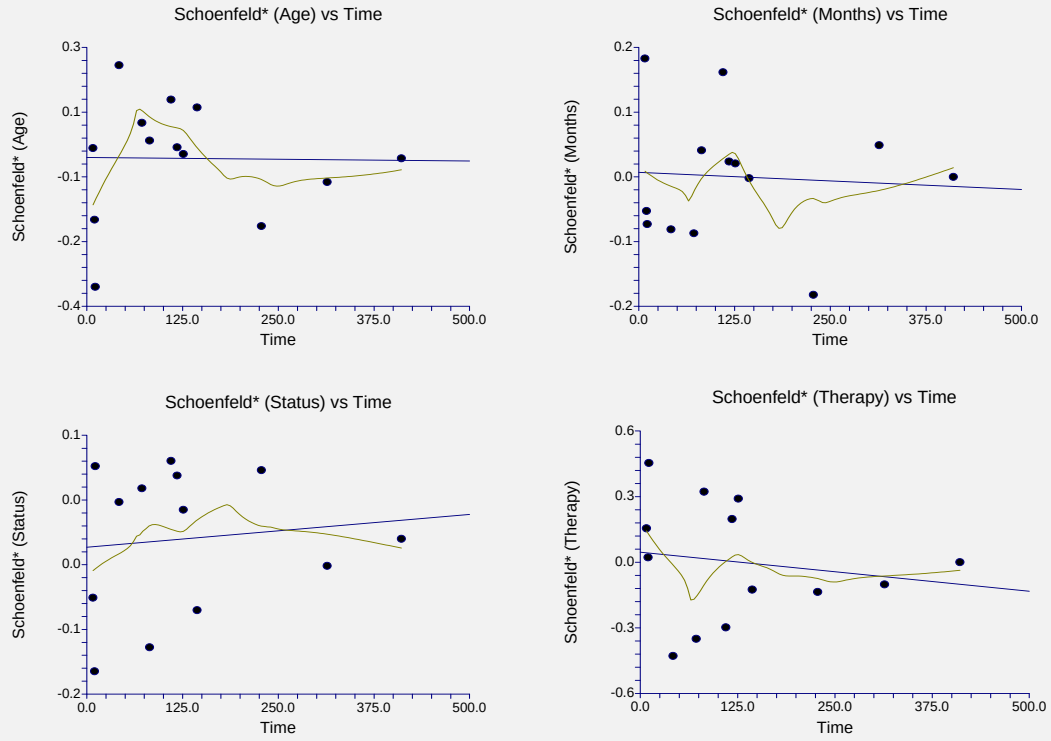


The Schoenfeld residuals are plotted for two reasons. First of all, these plots are useful in assessing whether the proportional hazards assumption is met. If the least squares line is horizontal and the lowess curve seems to track the least squares line fairly well, the proportional hazard assumption is reasonable.

Second, points that are very influential in determining the estimated regression coefficient for a covariate show up as outliers on these plots. When influential points are found, it is important to make sure that the data associated with these points are accurate. It is not advisable to remove these influential points unless a specific reason can be found for doing so.

Many authors suggest that the scaled Schoenfeld residuals are more useful than these, unscaled, residuals.

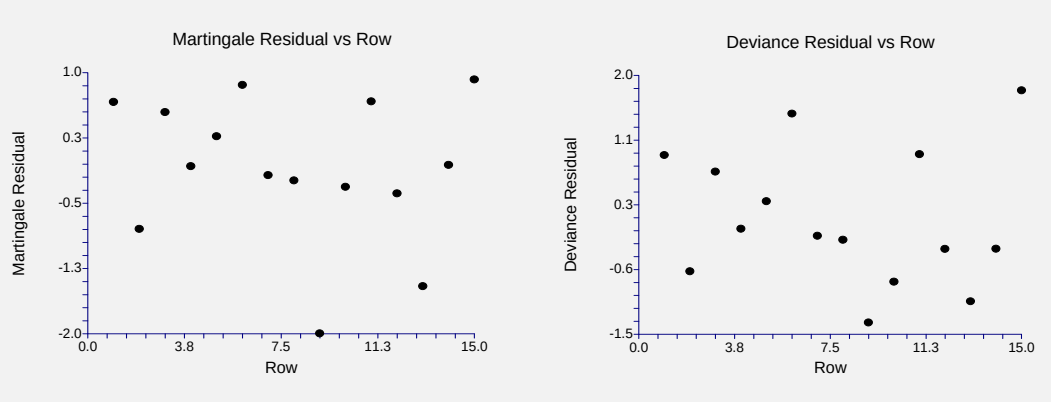
Plots of Scaled Schoenfeld Residuals Versus Time



The scaled Schoenfeld residuals are plotted for two reasons. First of all, these plots are useful in assessing whether the proportional hazards assumption is met. If the least squares line is horizontal and the lowess curve seems to track the least squares line fairly well, the proportional hazard assumption is reasonable.

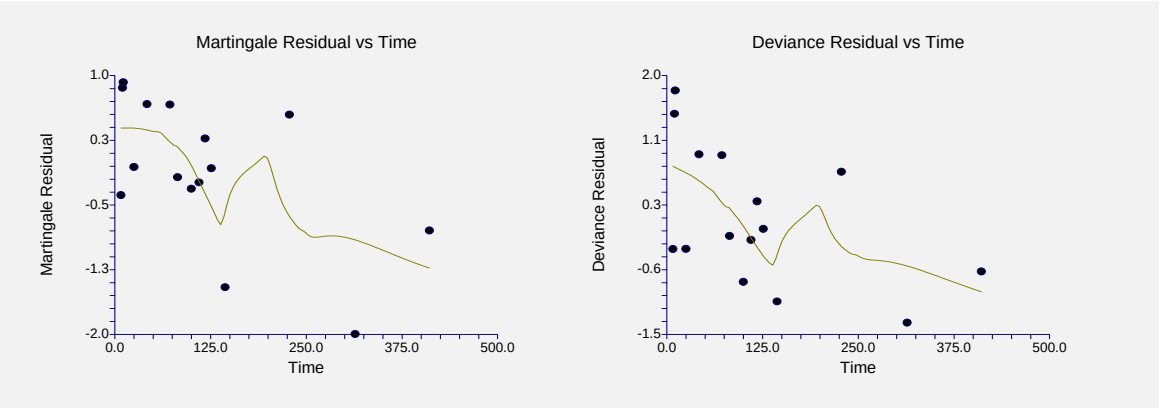
Second, points that are very influential in determining the estimated regression coefficient for a covariate show up as outliers on these plots. When influential points are found, it is important to make sure that the data associated with these points are accurate. It is not advisable to remove these influential points unless a specific reason can be found for doing so.

Plots of Residuals Versus Row



These plots are made to allow you to find outliers. These outliers should be double-checked to be certain that the data are not in error. You should not routinely remove outliers unless you can find a good reason for doing so. Often, the greatest insight during an investigation comes while considering why these outliers are different.

Plots of Residuals Versus Time



These plots are made to allow you to find outliers. These outliers should be double-checked to be certain that the data are not in error. You should not routinely remove outliers unless you can find a good reason for doing so. Often, the greatest insight during an investigation comes while considering why these outliers are different.

Example 2 – Subset Selection

This section presents an example of how to conduct a subset selection. We will again use the LUNGCANCER database that was used in Example 1. In this run, we will be trying to find a subset of the covariates that should be kept in the regression model.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Cox Regression window.

1 Open the LUNGCANCER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LungCancer.s0**.
- Click **Open**.

2 Open the Cox Regression window.

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Cox Regression**. The Cox Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will load the default template.

3 Specify the variables.

- On the Cox Regression window, select the **Variables tab**.
- Enter **Time** in the **Time Variable** box.
- Set the **Ties Method** to **Efron**.

- Enter **Censor** in the **Censor Variable** box.
- Enter **Status-Therapy** in the **X's: Numeric Independent Variables** box.

4 Specify the model.

- On the Cox Regression window, select the **Models tab**.
- Set the **Subset Selection** box to **Hierarchical Forward with Switching**.

5 Specify the reports.

- On the Cox Regression window, select the **Reports tab**.
- Check all of the reports. Although under normal circumstances you would not need all of the reports, we will view them all here so they can be annotated.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Subset Selection Summary Section

Number of Terms	Number of X's	Log Likelihood	R-Squared Value	R-Squared Change
0	0	-23.3349	0.0000	0.0000
1	1	-21.8803	0.1763	0.1763
2	2	-21.0354	0.2641	0.0878
3	3	-20.1352	0.3473	0.0832
4	4	-20.1143	0.3491	0.0018

This report shows the best log-likelihood value for each subset size. In this example, it appears that a model with three terms provides the best model. Note that adding the fourth variable does not increase the R-squared value very much.

No. Terms

The number of terms.

No. X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function evaluated at the maximum likelihood estimates. Our goal is to find a subset size above which little is gained by adding more variables.

R-Squared Value

This is the value of *R*-squared calculated using the formula

$$R_k^2 = 1 - \exp\left[\frac{2}{n}\left(L_0 - L_k\right)\right]$$

as discussed in the introduction. We are looking for the subset size after which this value does not increase by a meaningful amount.

R-Squared Change

This is the increase in *R*-squared that occurs when each new subset size is reached. Search for the subset size below which the *R*-squared value does not increase by more than 0.02 for small samples or 0.01 for large samples.

In this example, the optimum subset size appears to be three terms.

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	Log Likelihood	R-Squared	Term Entered	Terms Removed
1	Begin	0	0	-23.3349	0.0000		
2	Add	1	1	-21.8803	0.1763	MONTHS	
3	Add	2	2	-21.0354	0.2641	STATUS	
4	Add	3	3	-20.1352	0.3473	AGE	
5	Add	4	4	-20.1143	0.3491	THERAPY	

This report shows the highest log likelihood for each subset size. In this example, it appears that three terms provide the best model. Note that adding THERAPY does not increase the R-squared value very much.

Action

This item identifies the action that was taken at this step. A term was added, removed, or two were switched.

No. Terms

The number of terms.

No. X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function after the completion of this step. Our goal is to find a subset size above which little is gained by adding more variables.

R-Squared Value

This is the value of *R*-squared calculated using the formula

$$R_k^2 = 1 - \exp\left[\frac{2}{n}\left(L_0 - L_k\right)\right]$$

as discussed in the introduction. We are looking for the subset size after which this value does not increase by a meaningful amount.

Terms Entered and Removed

These columns identify the terms added, removed, or switched.

Discussion of Example 2

After considering these reports, it was decided to include AGE, MONTHS, and STATUS in the final regression model. Another run is performed using only these independent variables. A complete residual analysis would be necessary before the equation is finally adopted.

Regression Coefficients Section

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	Prob Level	Pseudo R2
Age	0.041940	0.033413	1.0428	60.3333	1.2552	0.2094	0.1361
Months	0.063724	0.032004	1.0658	12.6000	1.9912	0.0465	0.2838
Status	-0.031482	0.019680	0.9690	57.3333	-1.5997	0.1097	0.2037

This report displays the results of the proportional hazards estimation. Note that the Wald tests indicate that only MONTHS is statistically significant. Because of the small sample size of this example and because they add a great deal to the R-squared value, we have added AGE and STATUS to the final model.

Example 3 – Cox Regression with Categorical Variables

This example will demonstrate the analysis of categorical independent variables. A study was conducted to evaluate the influence on survival time of three variables: Age, Gender, and Treatment. The ages of the study participants were grouped into three age categories: 20, 40, and 60. The first age group (20) was selected as the reference group. The female group was selected as the reference group for Gender. The Treatment variable represented three groups: a control and two treatment groups. The control group was selected as the reference group for Treatment. The data for this study are contained in the COXREG database.

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Cox Regression window.

1 Open the COXREG dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **COXREG.s0**.
- Click **Open**.

2 Open the Cox Regression window.

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Cox Regression**. The Cox Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will load the default template.

3 Specify the variables.

- On the Cox Regression window, select the **Variables tab**.
- Enter **Time** in the **Time Variable** box.
- Set the **Ties Method** to **Efron**.
- Enter **Status** in the **Censor Variable** box.
- Enter **Count** in the **Frequency Variable** box.
- Enter **Treatment AGE GENDER** in the **X's: Categorical Independent Variables** box.

4 Specify the model.

- On the Cox Regression window, select the **Model tab**.
- Set **Which Model Terms** to **Up to 2-Way**.

5 Specify the reports.

- On the Cox Regression window, select the **Reports tab**.
- Check the **Run Summary**, **Regression Coefficients**, **C.L. of Regression Coefficients**, **Analysis of Deviance**, and **Log-Likelihood and R-Squared** reports.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Section

Run Summary Section

Parameter	Value	Parameter	Value
Rows Read	73	Time Variable	Time
Rows Filtered Out	0	Censor Variable	Status
Rows Missing X's	0	Frequency Variable	Count
Rows Processed	73	Subset Method	None
Rows Prediction Only	0	Ind. Var's Available	3
Rows Failed	54	No. of X's in Model	13
Rows Censored	19	Iterations	6
Sum of Frequencies	137	Maximum Iterations	20
Sum Censored Freqs	83	Convergence Criterion	1E-09
Sum Failed Freqs	54	Achieved Convergence	2.499645E-16
Final Log Likelihood	-222.9573	Completion Message	Normal completion

This report summarizes the characteristics of the dataset. Note that 137 individuals were included in this study of which 83 were censored.

Regression Coefficients Section

Regression Coefficients Section

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	Prob Level	Pseudo R2
B1 (Age=40)	0.178527	0.720145	1.1955	0.3066	0.2479	0.8042	0.0014
B2 (Age=60)	0.747377	0.652733	2.1115	0.3431	1.1450	0.2522	0.0283
B3 (Gender="M")	0.199655	0.629734	1.2210	0.5182	0.3170	0.7512	0.0022
B4 (Treatment="T1")	0.315769	0.707474	1.3713	0.3358	0.4463	0.6554	0.0044
B5 (Treatment="T2")	0.606087	0.674007	1.8332	0.3212	0.8992	0.3685	0.0177
B6 (Age=40)*(Gender="M")	0.420448	0.701899	1.5226	0.1679	0.5990	0.5492	0.0079
B7 (Age=60)*(Gender="M")	0.366048	0.709829	1.4420	0.1971	0.5157	0.6061	0.0059
B8 (Age=40)*(Treatment="T1")	-0.228646	0.872282	0.7956	0.1095	-0.2621	0.7932	0.0015
B9 (Age=40)*(Treatment="T2")	-0.442119	0.843234	0.6427	0.0876	-0.5243	0.6001	0.0061
B10 (Age=60)*(Treatment="T1")	-0.124997	0.851308	0.8825	0.1022	-0.1468	0.8833	0.0005

B11 (Age=60)*(Treatment="T2")							
	-1.726851	0.885161	0.1778	0.1168	-1.9509	0.0511	0.0780
B12 (Gender="M")*(Treatment="T1")							
	-0.976831	0.714997	0.3765	0.1752	-1.3662	0.1719	0.0398
B13 (Gender="M")*(Treatment="T2")							
	-0.553592	0.721736	0.5749	0.1606	-0.7670	0.4431	0.0129

This report displays the results of the proportional hazards estimation. Note that the names of the interaction terms are too long to fit in the space allotted, so the rest of the information appears on the next line.

Independent Variable

It is important to understand the variable names of the interaction terms. For example, consider the last name: (Gender=“M”)*(Treatment=“T2”). This variable was created by multiplying two indicator variables. The first indicator is “1” when the gender is “M” and “0” otherwise. The second indicator is “1” when the treatment is “T2” and “0” otherwise. This portion of the gender-by-treatment interaction is represented by the product of these two variables.

The rest of the definitions are the same as before and so they are not repeated here.

Confidence Limits Section

Confidence Limits Section

Independent Variable	Regression Coefficient (B)	Lower 95.0% Confidence Limit of B	Upper 95.0% Confidence Limit of B	Risk Ratio Exp(B)	Lower 95.0% C.L. of Exp(B)	Upper 95.0% C.L. of Exp(B)
B1 (Age=40)	0.178527	-1.232933	1.589986	1.1955	0.2914	4.9037
B2 (Age=60)	0.747377	-0.531956	2.026709	2.1115	0.5875	7.5891
B3 (Gender="M")	0.199655	-1.034602	1.433911	1.2210	0.3554	4.1951
B4 (Treatment="T1")						
	0.315769	-1.070855	1.702392	1.3713	0.3427	5.4871
B5 (Treatment="T2")						
	0.606087	-0.714943	1.927116	1.8332	0.4892	6.8697
B6 (Age=40)*(Gender="M")						
	0.420448	-0.955248	1.796145	1.5226	0.3847	6.0264
B7 (Age=60)*(Gender="M")						
	0.366048	-1.025192	1.757288	1.4420	0.3587	5.7967
B8 (Age=40)*(Treatment="T1")						
	-0.228646	-1.938287	1.480996	0.7956	0.1440	4.3973
B9 (Age=40)*(Treatment="T2")						
	-0.442119	-2.094827	1.210589	0.6427	0.1231	3.3555
B10 (Age=60)*(Treatment="T1")						
	-0.124997	-1.793530	1.543536	0.8825	0.1664	4.6811
B11 (Age=60)*(Treatment="T2")						
	-1.726851	-3.461735	0.008033	0.1778	0.0314	1.0081
B12 (Gender="M")*(Treatment="T1")						
	-0.976831	-2.378199	0.424538	0.3765	0.0927	1.5289
B13 (Gender="M")*(Treatment="T2")						
	-0.553592	-1.968169	0.860985	0.5749	0.1397	2.3655

This report provides the confidence intervals for the regression coefficients and the risk ratios. The confidence coefficient, in this example 95%, was specified on the Format tab. Note that the names of the interaction terms are too long to fit in the space allotted, so the rest of the information appears on the next line.

Independent Variable

It is important to understand the variable names of the interaction terms. For example, consider the last name: (Gender=“M”)*(Treatment=“T2”). This variable was created by multiplying two indicator variables. The first indicator is “1” when the gender is “M” and “0” otherwise. The

second indicator is “1” when the treatment is “T2” and “0” otherwise. This portion of the gender-by-treatment interaction is represented by the product of these two variables.

The rest of the definitions are the same as before and so they are not repeated here.

Analysis of Deviance Section

Term(s) Omitted	DF	-2 Log Likelihood	Increase From Model Deviance (Chi Square)	Prob Level
All Terms	13	454.5022	8.5876	0.8033
AGE	2	447.2661	1.3515	0.5088
GENDER	1	446.0147	0.1001	0.7517
TREATMENT	2	446.7191	0.8044	0.6688
AGE*GENDER	2	446.3421	0.4275	0.8076
AGE*TREATMENT	4	451.1965	5.2819	0.2596
GENDER*TREATMENT	2	447.7827	1.8681	0.3930
None(Model)	13	445.9146		

This report is the Cox regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms are statistically significant after adjusting for all other terms in the model.

The DF (degrees of freedom) column indicates the number of binary variables needed to represent each term. The chi-square test is used to test the significance of all binary variables associated with a particular term.

Log Likelihood & R-Squared Section

Term(s) Omitted	DF	Log Likelihood	R-Squared Of Remaining Term(s)	Reduction From Model R-Squared
All Terms	13	-227.2511	0.0000	0.0608
AGE	2	-223.6331	0.0514	0.0093
GENDER	1	-223.0074	0.0601	0.0007
TREATMENT	2	-223.3595	0.0552	0.0055
AGE*GENDER	2	-223.1711	0.0578	0.0029
AGE*TREATMENT	4	-225.5982	0.0238	0.0369
GENDER*TREATMENT	2	-223.8914	0.0479	0.0129
None(Model)	13	-222.9573	0.0608	0.0000

This report displays the Log Likelihood and R-Squared that is achieved when each term is omitted from the regression model. The DF (degrees of freedom) column indicates the number of binary variables needed to represent each term. The chi-square test is used to test the significance of all binary variables associated with a particular term.

Example 4 – Validation of Cox Regression using Collett (1994)

Collett (1994), pages 156 and 157, present a dataset giving the results of a small study about kidney dialysis. This dataset contains two independent variables: Age and Sex. These data are contained in the NCSS database called COLLETT157.

Collett (1994) gives the estimated regression coefficients as 0.030 for Age and -2.711 for Sex. The chi-square test for Sex is 6.445 and the chi-square test for Age is 1.320. The Cox-Snell residual for the first patient is 0.3286. The martingale residual for this patient is 0.6714. The deviance residual for this patient is 0.9398. The Schoenfeld residuals for this patient are -1.0850 and -.2416.

We will run these data through NCSS and check that we obtain the same values. You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Cox Regression window.

1 Open the COLLETT157 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **COLLETT157.s0**.
- Click **Open**.

2 Open the Cox Regression window.

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Cox Regression**. The Cox Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will load the default template.

3 Specify the variables.

- On the Cox Regression window, select the **Variables tab**.
- Enter **Time** in the **Time Variable** box.
- Set the **Ties Method** to **Efron**.
- Enter **Status** in the **Censor Variable** box.
- Enter **Age Sex** in the **X's: Numerical Independent Variables** box.

4 Specify the reports.

- On the Cox Regression window, select the **Reports tab**.
- Check the **Run Summary**, **Regression Coefficients**, **Analysis of Deviance**, **Residuals**, and **Schoenfeld Residuals** reports.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Validation Report

Regression Coefficients Section

Independent Variable	Regression Coefficient (B)	Standard Error of B	Risk Ratio Exp(B)	Mean	Wald Z-Value	Prob Level	Pseudo R2
B1 Age	0.030371	0.026237	1.0308	31.4615	1.1576	0.2470	0.1181
B2 Sex	-2.710762	1.095898	0.0665	1.7692	-2.4736	0.0134	0.3795

Analysis of Deviance Section

Term(s) Omitted	DF	-2 Log Likelihood	Increase From Model Deviance (Chi Square)	Prob Level
All Terms	2	40.9454	6.4779	0.0392
AGE	1	35.7880	1.3204	0.2505
SEX	1	40.9132	6.4456	0.0111
None(Model)	2	34.4676		

Residuals Section

Row	Time	Cox-Snell Residual		Martingale Residual		Deviance Residual	
1	8	0.3286	|.....	0.6714	||||.....	0.9398	||||.....
2	15	0.0785	|.....	0.9215	|||||||.....	1.8020	|||||||||||||
3	22	1.4331	|||||||.....	-0.4331	|.....	-0.3828	|.....
4	24	0.0939	|.....	0.9061	|||||||.....	1.7087	||||||||||||.
5	30	1.7736	|||||||||.....	-0.7736	|||||.....	-0.6334	||.....
6+	54	0.3117	|.....	-0.3117	|.....	-0.7895	||||.....
7	119	0.2655	|.....	0.7345	|||||.....	1.0877	|||||||.....
8	141	0.5386	||.....	0.4614	||.....	0.5611	|.....
9	185	1.6523	|||||||||.....	-0.6523	||||.....	-0.5480	|.....
10	292	1.4234	||||||||.....	-0.4234	|.....	-0.3751	|.....
11	402	1.4207	|||||||.....	-0.4207	|.....	-0.3730	|.....
12	447	2.3927	|||||||||||||	-1.3927	|||||||||||||	-1.0201	|||||.....
13	536	1.5640	||||||||.....	-0.5640	|||.....	-0.4832	|.....

Schoenfeld Residuals Section

Row	Time	Resid Age		Resid Sex	
1	8	-1.0850	|.....	-0.2416	|||||.....
2	15	14.4930	|||||||||||...	0.6644	|||||||||||||
3	22	3.1291	||.....	-0.3065	|||||.....
4	24	-10.2215	|||||||.....	0.4341	||||||||.....
5	30	-16.5882	||||||||||||..	-0.5504	|||||||||||...
7	119	-17.8286	|||||||||||||	0.0000	|.....
8	141	-7.6201	|||||.....	0.0000	|.....
9	185	17.0910	||||||||||||.	0.0000	|.....
10	292	10.2390	||||||||.....	0.0000	|.....
11	402	2.8575	||.....	0.0000	|.....
12	447	5.5338	||||.....	0.0000	|.....
13	536	0.0000	|.....	0.0000	|.....

You can verify that are results matched those of Collett (1994) within rounding.

Chapter 566

Parametric Survival (Weibull) Regression

Introduction

This module fits the regression relationship between a positive-valued dependent variable (often time to failure) and one or more independent variables. The distribution of the residuals (errors) is assumed to follow the exponential, extreme value, logistic, log-logistic, lognormal, lognormal10, normal, or Weibull distribution. The data may include failed, left censored, right censored, and interval observations. This type of data often arises in the area of *accelerated life testing*.

When testing highly reliable components at normal stress levels, it may be difficult to obtain a reasonable amount of failure data in a short period of time. For this reason, tests are conducted at higher than expected stress levels. The models that predict failure rates at normal stress levels from test data on items that fail at high stress levels are called *acceleration models.*

The basic assumption of acceleration models is that failures happen faster at higher stress levels. That is, the failure mechanism is the same, but the time scale has been changed (shortened).

Technical Details

The linear regression equation is

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + Se$$

Here, S represents the value of a constant standard deviation, Y is a transformation of time (either *ln(t), log(t),* or just t), the X*'s* are one or more independent variables, the B's are the regression coefficients, and e is the residual (error) that is assumed to follow a particular probability distribution. The problem reduces to estimating the B's and S. The density functions of the eight distributions that are fit by this module were given in the Distribution Fitting section and will not be repeated here.

So that you can get the general idea, we will give detailed results for the lognormal distribution. The results for other distributions follow a similar pattern.

The lognormal probability density function may be written as

$$f(t \mid M, S) = \frac{1}{tS\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(t)-M}{S}\right)^2}$$

If we replace the location parameter, *M*, with the regression model, the density now becomes

$$f(t \mid B_0 \cdots B_p, S) = \frac{1}{tS\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(t) - \sum_{i=0}^{p} B_i X_i}{S}\right)^2\right\}$$

Maximum likelihood estimation consists of finding the values of the distribution parameters that maximize the log-likelihood of the data values. Loosely speaking, these are the values of the parameters which maximize the probability that the current set of data values occur.

The general form of the log-likelihood function is given by

$$L(\underline{P}) = \sum_{F} \ln\left(f(\underline{P}, t_k)\right) + \sum_{R} \ln\left(S(\underline{P}, t_k)\right) + \sum_{L} \ln\left(F(\underline{P}, t_k)\right) + \sum_{I} \ln\left(f(\underline{P}, t_{uk}) - f(\underline{P}, t_{lk})\right)$$

where *F* represents the set of failed items, *R* represents the set of right censored items, *L* represents the set of left censored items, and *I* represents the set of interval censored items. In the case of interval censored observations, t_{lk} represents the first time of the interval and t_{uk} represents the last time of the interval. Also, $\underline{P}$ represents the parameters, including *S* and the *B's*.

We employ the Newton-Raphson algorithm with numerical differentiation to obtain the maximum likelihood estimates. These estimates have been shown to have optimality characteristics in large samples (number of failures greater than 20). They have been shown to be competitive estimates even for sample sizes less than 20.

Data Structure

Survival data are somewhat more difficult to enter because of the presence of various types of censoring.

Time Variable(s)

One (or two in the case of interval data) variable is needed to contain the time values.

Censor Variable

Another variable is needed to indicate the type of censoring.

Failed or Complete

A failed observation is one in which the time until the terminal event was measured exactly; for example, the machine stopped working or the mouse died of the disease being studied.

Right Censored

A right censored observation provides a lower bound for the actual failure time. All that is known is that the failure occurred (or will occur) at some point after the given time value. Right censored observations occur when a study is terminated before all items have failed. They also occur when an item fails due to an event other than the one of interest.

Left Censored

A left censored observation provides an upper bound for the actual failure time. All we know is that the failure occurred at some point before the time value. Left censoring occurs when the items are not checked for failure until some time after the study has begun. When a failed item is found, we do not know exactly when it failed, only that it was at some point before the left censor time.

Interval Censored or Readout

An interval censored observation is one in which we know that the failure occurred between two time values, but we do not know exactly when. This type of data is often called *readout* data. It occurs in situations where items are checked periodically for failures.

Independent Variable(s)

One or more independent variables must be supplied also.

Sample Data

The following data, found in Nelson (1990), are quoted in many books and articles on accelerated testing. These data come from a temperature-accelerated life test of a Class-B insulation for electric motors. Ten motorettes were tested at each of four temperatures. When the testing was stopped, the following failure times were recorded. These data are stored in the MOTORS.S0 database.

MOTORS dataset

Hours	Censor	Count	Temperature
	1	1	130
8064	0	10	150
1764	1	1	170
2772	1	1	170
3444	1	1	170
3542	1	1	170
3780	1	1	170
4860	1	1	170
5196	1	1	170
5448	0	3	170
408	1	2	190
1344	1	2	190
1440	1	1	190
1680	0	5	190
408	1	2	220
504	1	3	220
528	0	5	220

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the probability distribution that is fit and the variables used in the analysis.

Time Variables

Y: Time Variable

This variable contains the time values for both failed and censored observations. When interval (readout) data are used, this variable specifies the ending time of the interval.

Negative time values are treated as missing values. Zero time values are not allowed. They may be automatically replaced by the value in the Zero field.

These time values represent elapsed times. If your dataset is made up of dates (such as the failure date), you must subtract the starting date from the failure date so that you can analyze the elapsed time.

Start Time Variable

This variable contains the starting time for interval (readout) data. Hence its value is only used when the row's censor value indicates an interval data type.

Negative time values are treated as missing values. Rows with zero starting time values are reset to left censored type data.

Zero Time Replacement Value

Under normal conditions, a respondent beginning the study is "alive" and cannot "die" until after some small period of time has elapsed. Hence, a time value of zero is not defined and is ignored (treated as a missing value). If a zero time value does occur, it is replaced by this positive amount. If you do not want zero time values replaced, enter a "0.0" here.

This option is used when a "zero" on the database does not actually mean zero time. Instead, it means that the response occurred before the first reading was made and so the actual survival time is only known to be less.

Censor Variable

Censor Variable

The values in this optional variable indicate the type of censoring active for each row. Four possible data types may be entered: failed (complete), right censored, left censored, or interval. The values used to indicate each data type are specified in the four boxes to the right. These values may be text or numeric.

Failed

When this value is entered in the Censor Variable, the corresponding time value is treated as a failed observation. The value may be a number or a letter. We suggest the letter "F" when you are in doubt as to what to use.

A failed observation is one in which the time until the event of interest was measured exactly; for example, the machine stopped working or the mouse died of the disease being studied. The exact failure time is known.

Right

When this value is entered in the Censor Variable, the corresponding time value is treated as a right censored data value. The value may be a number or a letter. We suggest the letter "R" when you are in doubt as to what to use.

A right censored observation provides a lower bound for the actual failure time. All that is known is that the failure time occurred (or will occur) at some point after the given time value. Right censored observations occur when a study is terminated before all items have failed. They also occur when an item fails due to an event other than the one of interest.

Left

When this value is entered in the Censor Variable, the corresponding time value is treated as a left censored data value. The value may be a number or a letter. We suggest the letter "L" when you are in doubt as to what to use.

A left censored observation provides an upper bound for the actual failure time. All we know is that the failure time occurred at some point before the time value. Left censoring occurs when the items are not checked until some time after the study has begun. When a failed item is found, we do not know exactly when it failed, only that it was at some point before the left censor time.

Interval

When this value is entered in the Censor Variable, the corresponding time value is treated as an interval censored data value. The value may be a number or a letter. We suggest the letter "I" when you are in doubt as to what to use. When interval censoring is specified, the program uses both the Time Variable and the Start Time Variable.

An interval censored observation is one in which we know that the failure occurred between the two time values, but we do not know exactly when. This type of data is often called *readout* data. It occurs in situations where items are checked periodically for failures.

Note that when interval data are obtained, the first observation is usually left censored and the last observation is usually right censored.

Independent Variables

X's: Independent Variables

Specify additional independent variables. You can leave this option blank or you can leave the Stress Variable blank, but you cannot leave both blank.

These variables may be thought of as additional variables for which statistical adjustment is desired. They can contain both discrete and continuous variables. If discrete variables are specified, it is up to you to specify the appropriate number of dummy variables. For example, suppose you have three suppliers. Since this has three possible values, two indicator variables will be needed to specify the appropriate information.

Frequency Variable

Frequency Variable

This variable gives the count, or frequency, of the time displayed on that row. When omitted, each row receives a frequency of one. Frequency values should be positive integers. This is usually used to indicate the number of right censored values at the end of a study or the number of failures occurring within an interval. It may also be used to indicate ties for failure data.

Stress Variable

Stress Variable

This variable contains the values of the independent variable that will be transformed according to either the Arrhenius or Power transformation. This variable is optional, although when it is not specified, several of the reports and graphs will not be displayed.

You can leave this option blank or you can leave the Independent Variables option blank, but you cannot leave both blank.

Stress A

Specify the value of *A* in the Arrhenius transformation *X=A/(Stress+B).* If *A* is zero, the logarithm of stress is used and the model is X=log(stress).

Set *A* to 1000 for Arrhenius model. Set *A* to 0 for Power model.

Stress B

Specify *B* in the transformation *X=A/(Stress+B).* Ignore it by setting *B* to zero.

For the Arrhenius model, *B* is set to 273.16 when stress is measured in degrees Celsius. Usually, *A* and *B* are set to convert temperature values to degrees Kelvin.

For the Power acceleration model, this value is ignored.

Probability Distribution

Distribution

This option specifies which probability distribution is fit. All results are for the specified probability distribution. Usually, you will use one of the distributions that is based on the logarithm of time such as the lognormal, Weibull, exponential, or log-logistic.

Alpha Level

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Estimation Tab

The following options control the searching algorithms used during parameter estimation.

Estimation Options

Maximum Iterations

Many of the parameter estimation algorithms are iterative. This option assigns a maximum to the number of iterations used in any one algorithm. We suggest a value of at least 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained. If the number of iterations reaches this amount, you should re-run your analysis with a larger value.

Minimum Relative Change

This value is used to control the iterative algorithms used in maximum likelihood estimation. When the relative change in all of the parameters is less than this amount, the iterative procedure is terminated.

Parameter Adjustment

Newton's method calculates a change for each parameter value at each step. Instead of taking the whole parameter change, this option lets you take only a fraction of the indicated change. For datasets that diverge, taking only partial steps may allow the algorithm to converge. In essence, the algorithm tends to over correct the parameter values. This factor allows you to dampen this over correction. We suggest a value of about 0.2. This may increase the number of iterations (and you will have to increase the Maximum Iterations accordingly), but it provides a greater likelihood that the algorithm will converge.

Starting Sigma

Specify a starting value for *S*, the standard deviation of the residuals (errors). Select '0 - Data' to calculate an appropriate value from the data. If convergence fails, try a different value.

Derivatives

This value specifies the machine precision value used in calculating numerical derivatives. Slight adjustments to this value can change the accuracy of the numerical derivatives (which impacts the variance/covariance matrix estimation).

Remember from calculus that the derivative is the slope calculated at a point along the function. It is the limit found by calculating the slope between two points on the function curve that are very close together. Numerical differentiation mimics this limit by calculating the slope between two function points that are very close together and then computing the slope. This value controls how close together these two function points are.

Numerical analysis suggests that this distance should be proportional to the machine precision of the computer. We have found that our algorithm achieves four-place accuracy in the variance-covariance matrix no matter what value is selected here (within reason). However, increasing or decreasing this value by two orders of magnitude may achieve six or seven place accuracy in the variance-covariance matrix. We have found no way to find the optimal value except trial and error.

Note that the parameter estimates do not seem to be influenced a great deal, only their standard errors.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report ... Residual Report

Each of these options specifies whether the indicated report is calculated and displayed.

Percent Failing and Failure Time Percentile Reports

These options indicate whether to display the corresponding report and which rows are to be shown. Usually, you will add rows with missing time values at the end of the database to be reported on. The percent failing will then be estimated for those values.

You can choose to omit these reports entirely, show only those rows with missing time values, or show the results for all rows.

Select Plots

Stress - Time Plot ... X - Residual Plots

Each of these options specifies whether the indicated plot is displayed.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also, note that all calculations are performed in double precision regardless of which option you select here. Single precision is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Percent Failing Report Calculation Times

This option specifies a list of times at which the percent failing is reported on the Percent Failing Section report. Individual values are separated by commas. You can specify a sequence by specifying the minimum and maximum separated by a colon and putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers. Use ‘(10)’ alone to specify ten values between zero and the maximum value found in the data.

Note that each time is used for all selected observations.

Failure Time Report and Stress-Time Plot Percentiles

This option specifies a list of percentiles (range 1 to 99) at which the failure time is reported, one percentile per line on the Failure Time Percentile report. It also specifies which percentiles are shown on the Stress plot. The values should be separated by commas.

You can specify sequences with a colon, putting the increment inside parentheses after the maximum in the sequence. For example: 5:25(5) means 5,10,15,20,25 and 1:5(2),10:20(2) means 1,3,5,10,12,14,16,18,20. Note that this option also controls which percentiles are displayed on the Stress - Time plot.

Report Options – Decimal Places

Time ... Stress Decimals

This option specifies the number of decimal places shown on reported time, probability, and stress values.

Plot Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the appropriate legend name.

Stress - Time Plot to X - Resid Plots Tabs

These options control the attributes of the plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Y Log Scale (Vertical Axis of Stress - Time Plot Only)

This box lets you designate whether to display the vertical axis in a regular or logarithmic scale. This option is not appropriate when you have used logarithms to the base e.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Min and Max Stress

These option let you specify the minimum and maximum stress values displayed on the Stress - Time Plot. Since you are usually interested in values of stress lower than in your data, you will want to set these values carefully so that you include values at the typical stress levels.

Number Stresses or Number Predicted

This options sets resolution of the plot along the horizontal axis. A value near 50 is usually adequate.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines & Symbols Tab

These options specify the attributes of the lines and symbols used to display the percentiles in the Stress - Time plot.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Plotting Symbols

Right Censored – Predicted

This option specifies the symbol used for each type of data, censored, failed, and predicted. These symbols are provided to allow the various censoring types to be identified, even on black and white printers.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Lognormal Regression

This section presents an example of how to fit a lognormal regression. The data used were shown above and are found in the MOTORS database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Parametric Survival (Weibull) Regression window.

1 Open the MOTORS dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MOTORS.S0**.
- Click **Open**.

2 Open the Distribution Regression window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Parametric Survival Regression**. The Parametric Survival (Weibull) Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Parametric Survival (Weibull) Regression window, select the **Variables tab**.
- Double-click in the **Y: Time Variable** box. This will bring up the variable selection window.
- Select **HOURS** from the list of variables and then click **Ok**.
- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **CENSOR** from the list of variables and then click **Ok**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **COUNT** from the list of variables and then click **Ok**.
- Double-click in the **Stress Variable** box. This will bring up the variable selection window.
- Select **TEMP** from the list of variables and then click **Ok**. Note that the default values of Stress A and Stress B are appropriate for this problem.
- Set the **Distribution** to **Log10normal**.

4 Specify the reports.

- On the Parametric Survival (Weibull) Regression window, select the **Reports tab**.
- Enter **10000:100000(10000)** in the **Percent Failing Report Calculation Times** box.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Hours Minimum	Hours Maximum
Missing or Prediction	1			
Failed	12	17	408.0	5196.0
Right Censored	4	23	528.0	8064.0
Left Censored	0	0		
Interval Censored	0	0		
Total (Nonmissing)	16	40	408.0	8064.0

Means for Rows with Failures

Variable	Mean
Hours	1919.412
Temp	190.5882

This report displays a summary of the data that were analyzed. Scan this report to determine if there are any obvious data-entry errors by double-checking the counts and the minimum and maximum.

The means given for each variable are for the noncensored rows.

Parameter Estimation Section

Maximum Likelihood Parameter Estimation Section

Parameter Name	Parameter Estimate	Standard Error	Z Value	Prob Level	Lower 95.0% C.L.	Upper 95.0% C.L.
Intercept	-6.018403	0.9464583	-6.3589	0.000000	-7.873427	-4.163379
Temp	4.31048	0.4364686	9.8758	0.000000	3.455017	5.165942
Sigma	0.2591772	4.730405E-02	5.4790	0.000000	0.1812336	0.3706422

R-Squared	0.531405
Log Likelihood	-148.5373
Iterations	55

This report displays parameter estimates along with standard errors, significance tests, and confidence limits. Note that the significance levels and confidence limits all use large sample formulas. How large is a large sample? We suggest that you only use these results when the number of failed items is greater than twenty.

Parameter Estimates

These are the maximum likelihood estimates (MLE) of the parameters. They are the estimates that maximize the likelihood function. Details are found in Nelson (1990) pages 287 - 295.

Standard Error

The standard errors are the square roots of the diagonal elements of the estimated Variance Covariance matrix.

Z Value

The z value is equal to the parameter estimate divided by the estimated standard error. This ratio, for large samples, follows the normal distribution. It is used to test the hypothesis that the parameter value is zero. This value corresponds to the t value that is used in multiple regression.

Prob Level

This is the two-tailed p-value for testing the significance of the corresponding parameter. You would deem independent variables with small p-values (less than 0.05) important in the regression equation.

Upper and Lower 100(1-Alpha)% Confidence Limits

These are the lower and upper confidence limits for the corresponding parameters. They are large sample limits. They should be ignored when the number of failed items is less than thirty. For the regression coefficients *B*, the formulas are

$$CL_i = \hat{B}_i \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{B}_i} \qquad i = 0,\cdots,p$$

where $\hat{B}_i$ is the estimated regression coefficient, $\hat{\sigma}_{\hat{B}_i}$ is its standard error, and *z* is found from tables of the standard normal distribution.

For the estimate of sigma, the formula is

$$CL = \hat{S}\exp\left\{\frac{\pm z_{1-\alpha/2}\hat{\sigma}_{\hat{S}}}{\hat{S}}\right\}$$

R-Squared

R-Squared reflects the percent of variation in log(time) explained by the independent variables in the model. A value near zero indicates a complete lack of fit, while a value near one indicates a perfect fit.

Note that this R-Squared value is computed for the failed observations only. Censored observations are ignored.

Log Likelihood

This is the value of the log likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Iterations

This is the number of iterations that were required to solve the likelihood equations. If this is greater than the maximum you specified, you will receive a warning message. You should then increase the Maximum Iterations and rerun the analysis.

Variance Covariance Matrix

Inverse of Fisher Information Matrix

	Intercept	Temp	Sigma
Intercept	0.8957834	-0.4123897	-8.21986E-03
Temp	-0.4123897	0.1905048	4.377223E-03
Sigma	-8.21986E-03	4.377223E-03	2.237673E-03

This table gives an estimate of the asymptotic variance covariance matrix which is the inverse of the Fisher information matrix. The elements of the Fisher information matrix are calculated using numerical differentiation.

Percent Failing Section

Percent Failing Section

Row	Temp	Hours	Percent Failing	Lower 95.0% C.L.	Upper 95.0% C.L.
1	130.0	10000.0	0.4689	0.0159	12.2509
1	130.0	20000.0	7.5434	0.9597	40.7218
1	130.0	30000.0	22.4510	4.5245	63.8811
1	130.0	40000.0	39.1662	9.9947	78.8708
1	130.0	50000.0	53.9401	15.9924	87.8111
1	130.0	60000.0	65.7053	21.7004	92.9798
1	130.0	70000.0	74.6251	26.7964	95.9395
1	130.0	80000.0	81.2324	31.2200	97.6345
1	130.0	90000.0	86.0786	35.0227	98.6098
1	130.0	100000.0	89.6239	38.2932	99.1751

This report displays the estimated percent failing at the time values that were specified in the Report Times box of the Reports Tab for each observation with a missing time value. In our example, the first row of the Motors database is missing. The value of Temp (the stress variable) equal to 130 degrees. Reliability is one minus probability of failure. Thus, the reliability at 80,000 hours at a temperature of 130 degrees is 100-81.2324 which is 18.7676%. The confidence limits for reliability may also be converted from the percent failing confidence limits by subtracting from 100.

Percent Failing

The percent failing at a particular temperature is calculated as

$$100 \times \hat{F}(t \mid X_i) = 100 \times \hat{F} = 100 \times F\left(\frac{\log(t) - \sum_{i=0}^{p} x_{ki} \hat{B}_i}{\hat{S}} \right)$$

where *F(z)* is the cumulative distribution of *f(z)*, the probability density function. That is,

$$F(z) = \int_0^z f(t|B_0, B_1, S, X)dt$$

Confidence Limits for Percent Failing

The confidence limits for this estimate are computed using the following formulas from Nelson (1990) page 296. Note that these estimates are large sample estimates based on the assumption

that the distribution of *F* is asymptotically normal. We recommend that the number of failures be at least thirty when using these estimates.

$$\hat{F}_{lower}(t|X_i) = \frac{\hat{F}}{\hat{F} + (1-\hat{F})\exp\left\{\dfrac{z_{1-\alpha/2}\sigma_{\hat{F}}}{\hat{F}\left(1-\hat{F}\right)}\right\}}$$

$$\hat{F}_{upper}(t|X_i) = \frac{\hat{F}}{\hat{F} + (1-\hat{F})\exp\left\{\dfrac{-z_{1-\alpha/2}\sigma_{\hat{F}}}{\hat{F}\left(1-\hat{F}\right)}\right\}}$$

where

$$\sigma_{\hat{F}} = \sqrt{\sum_{i=0}^{p+1}\sum_{j=0}^{p+1} h_i h_j vc_{ij}}$$

$$h_i = \frac{-x_i g\left(\dfrac{y(t) - \sum x_i \hat{B}_i}{\hat{S}}\right)}{\hat{S}} \qquad i = 0, \cdots, p$$

$$h_{p+1} = -\frac{y(t) - \sum x_i \hat{B}_i}{\hat{S}^2}$$

and vc_{ij} is the corresponding element from the variance covariance matrix. The function *y(t)* is *ln(t)* for the Weibull, log-logistic, exponential, and lognormal distributions, *log(t)* for the lognormal10 distribution, and simply *t* for the normal, extreme value, and logistic distributions. The value of *g(x)* depends on the distribution. For the Weibull, exponential, and extreme value distributions

$$g(z) = e^{z - e^z}$$

For the normal, lognormal, and lognormal10 distributions

$$g(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$$

For the logistic and log-logistic distributions

$$g(z) = \frac{e^z}{\left(1+e^z\right)^2}$$

Failure Time Percentile Section

Failure Time Percentile Section

Row	Temp	Percentile	Estimated Hours	Lower 95.0% C.L.	Upper 95.0% C.L.
1	130.0	10.0000	21937.2	11781.4	40847.6
1	130.0	50.0000	47133.6	24111.6	92137.3
1	130.0	90.0000	101269.8	44892.1	228449.2

This report displays failure time percentiles and confidence intervals for those percentiles specified in the Report Percentiles box of the Report tab. For example, the median failure time is 47,135.1 hours. The 95% confidence limits for the median time are 24,106.6 to 92,162.2 hours.

The confidence limits rely on the asymptotic normality of the distribution of the percentiles. The sample size should be greater than thirty failed items before you use these confidence limits. The formulas for these limits are given in Nelson (1990) page 295.

Percentile

This is the percentile being found. For example, the value of 50 here refers to the median failure time.

Estimated Hours

The estimated time value (dependent variable) at which *100P* of the items are expected to fail. The percentile is found by solving the equation

$$P = F\left(\frac{y(t) - \sum_{i=0}^{p} x_{ki}\hat{B}_i}{\hat{S}}\right)$$

for *y(t)*. The function *y(t)* is *ln(t)* for the Weibull, log-logistic, exponential, and lognormal distributions, *log(t)* for the lognormal10 distribution, and simply *t* for the normal, extreme value, and logistic distributions. *F(t)* is the cumulative distribution function.

Confidence Limits for a Percentile

The confidence limits are computed as follows. First compute

$$u_p = F^{-1}(P)$$

Next compute

$$\sigma_{\hat{t}_p} = \sqrt{\sum_{i=0}^{p+1}\sum_{j=0}^{p+1} x_i x_j vc_{ij}}$$

where vc_{ij} is the corresponding element of the variance covariance matrix and

$$x_0 = 1$$
$$x_1 = X_1$$
$$\vdots$$
$$x_p = X_p$$
$$x_{p+1} = u_P$$

Finally, for the lognormal, Weibull, exponential, and log-logistic distributions, compute

$$\hat{t}_{lower,p} = e^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} - z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

$$\hat{t}_{upper,p} = e^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} + z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

For the lognormal10 distribution, compute

$$\hat{t}_{lower,p} = 10^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} - z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

$$\hat{t}_{upper,p} = 10^{\left(\sum_{i=0}^{p} x_i B_i + u_p \hat{S} + z_{1-\alpha/2}\sigma_{\hat{t}_p}\right)}$$

For the normal, extreme value, and logistic distributions, compute

$$\hat{t}_{lower,p} = \sum_{i=0}^{p} x_i B_i + u_p \hat{S} - z_{1-\alpha/2}\sigma_{\hat{t}_p}$$

$$\hat{t}_{upper,p} = \sum_{i=0}^{p} x_i B_i + u_p \hat{S} + z_{1-\alpha/2}\sigma_{\hat{t}_p}$$

Residual Section

Residual Section

Row	(T) Hours	Log10(T)	Predicted Log10(T)	Raw Residual	Standardized Residual	Cox-Snell Residual
1			4.673331			
2R	8064.0	3.90655	4.168003	-0.2614523	-1.008778	0.1702434
3	1764.0	3.246499	3.708286	-0.4617874	-1.781744	3.811264E-02
4	2772.0	3.442793	3.708286	-0.2654927	-1.024368	0.1658548
5	3444.0	3.537063	3.708286	-0.1712228	-0.6606401	0.293595
6	3542.0	3.549248	3.708286	-0.1590374	-0.6136243	0.3143435
7	3780.0	3.577492	3.708286	-0.1307942	-0.5046515	0.3665836
8	4860.0	3.686636	3.708286	-2.164969E-02	-8.353239E-02	0.6286976
9	5196.0	3.715669	3.708286	7.383185E-03	2.848702E-02	0.7161357
10R	5448.0	3.736237	3.708286	2.795114E-02	0.1078457	0.7829427
11	408.0	2.61066	3.288272	-0.6776116	-2.614473	4.47828E-03
12	1344.0	3.128399	3.288272	-0.1598725	-0.6168463	0.3128878
13	1440.0	3.158362	3.288272	-0.1299092	-0.5012372	0.3683169
14R	1680.0	3.225309	3.288272	-6.296246E-02	-0.2429321	0.5175633
15	408.0	2.61066	2.722126	-0.1114662	-0.4300773	0.4058197
16	504.0	2.70243	2.722126	-1.969583E-02	-7.599371E-02	0.6343351
17R	528.0	2.722634	2.722126	5.075522E-04	1.958322E-03	0.6947109

This report displays the predicted value and residual for each row. If the analysis is being run on logarithms of time, all values are in logarithms. The report provides predicted values for all rows with values for the independent variables. Hence, you can add rows of data with missing time values to the bottom of your database and obtain the predicted values for them from this report. The report also allows you to obtain predicted values for censored observations.

You should ignore the residuals for censored observations, since the residual is calculated as if the time value was a failure.

Row

This is the number of the observation being reported on. Censored observations have a letter appended to the row number.

(T) Hours

This is the original value of the dependent variable.

Log10(T)

This is the transformed value of the dependent variable.

Predicted Log10(T)

This is the predicted transformed value of the dependent variable (usually time). Note that *y* depends on the distribution being fit. For the Weibull, exponential, lognormal, and log-logistic distributions, the *y* is *ln(t)*. For the lognormal10 distribution, *y* is *log(t)*. For the extreme value, normal, and logistic distributions, *y* is *t*. The formula for *y* is

$$\hat{y} = \sum_{i=0}^{p} x_i B_i$$

Raw Residual

This is the residual in the *y* scale. The formula is

$$r_k = y_k - \sum_{i=0}^{p} x_i B_i$$

Note that the residuals of censored observations are not directly interpretable, since there is no obvious value of *y*. The row is displayed so that you can see the predicted value for this censored observation.

Standardized Residual

This is the residual standardized by dividing by the standard deviation. The formula is

$$r_k' = \frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}$$

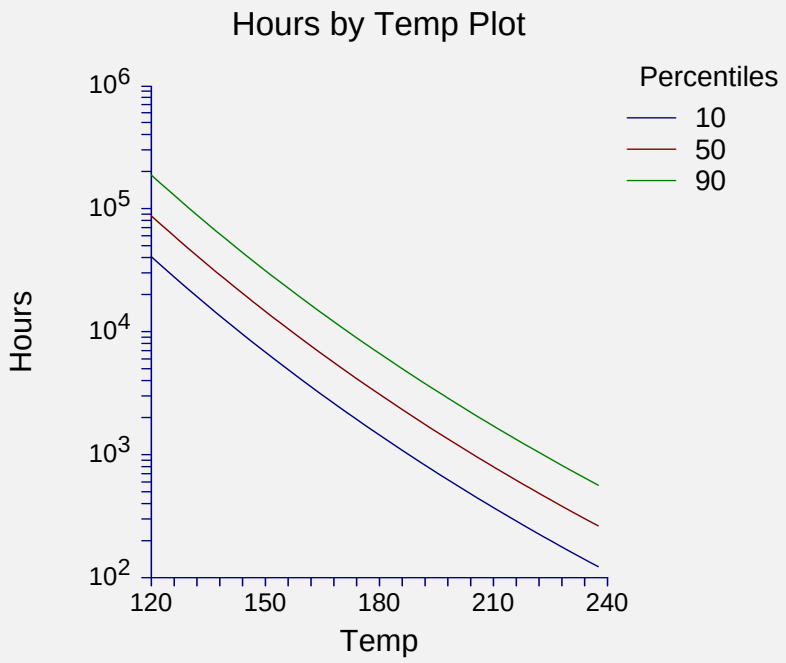
Cox-Snell Residual

The Cox-Snell residual is defined as

$$r_k'' = -\log\left\{1 - F\left(\frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}\right)\right\}$$

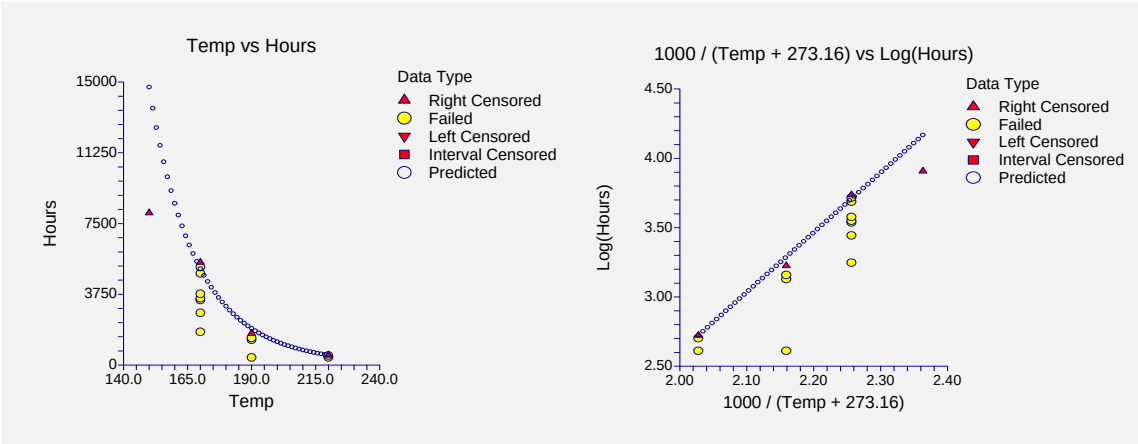
Here again, the residual does not have a direct interpretation for censored values.

Stress Plot



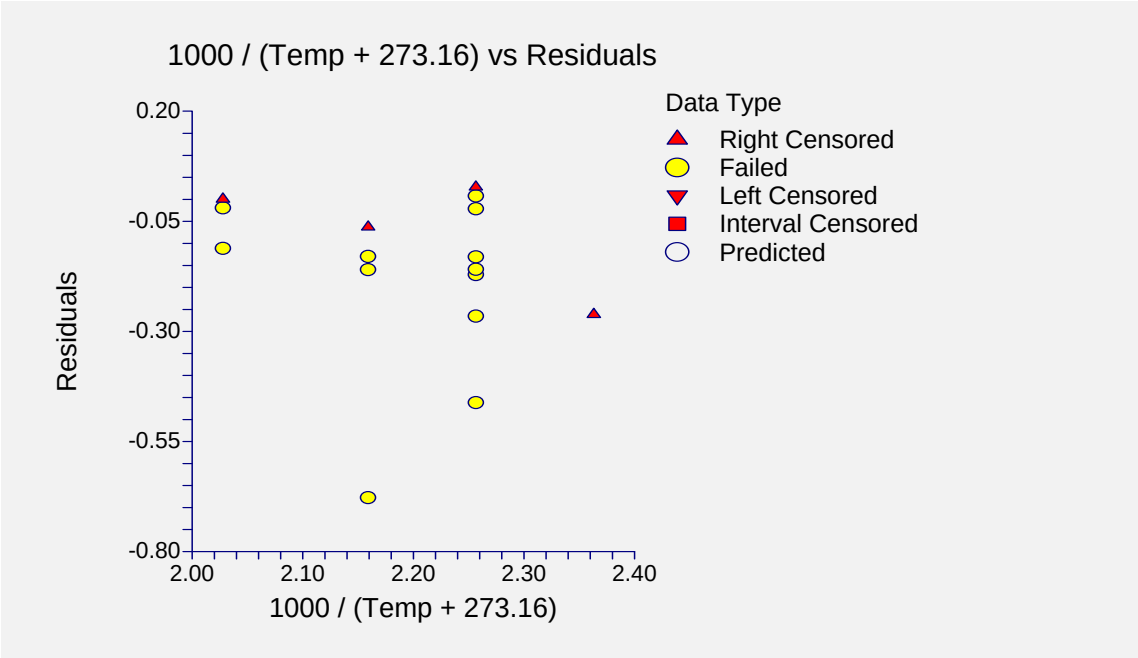
This plot displays the time on the vertical axis and the stress variable on the horizontal axis. The plotted lines represent the percentiles specified on the Reports tab window. This allows you to quickly view the percentiles for a wide range of stress values.

X-Y Plots



These plots show the data values from which the analysis was run. The plot on the left shows time versus stress in the original scale. The plot on the right shows time versus stress in the transformed metric. The prediction equation is also show on the chart. This lets you decide whether predictions are accurate. It also lets you study the goodness of fit.

Residual Plots



This plot shows the residuals in the transformed scale. You would study this chart just as you would any other residual versus independent variable plot from a multiple regression analysis. You are especially interested in finding outliers, as they can distort your results.

Discussion of Example

This example will look at an analysis of the electric motor data that was presented above and for which all of the above sample reports were generated. As mentioned earlier, a temperature-accelerated life test of a Class-B insulation was conducted using ten motors tested at each of four temperatures. When the testing was stopped, the failure times were recorded. These data are stored in the MOTORS.S0 database.

The purpose of this study was to determine the reliability of these motors at the normal operating temperature of 130°C by testing the reliability at higher temperatures. Note that at 150°C, none of the motors failed during the duration of the test.

The first step in the analysis is to determine if the fit is adequate. We look at the plots, the value of R-Squared, and the estimated value of sigma to determine this. The plots do not show any alarming points, although the residual plots show what might be a mild outlier in the 190°C batch.

Once the adequacy of the fit has been substantiated, we look at the Failure Time Percentile Section. This report provides the 10^{th}, 50^{th}, and 90^{th} percentiles. The estimated failure times for these percentiles are 21,937 hours, 47,134 hours, and 101,270 hours. That is, we would expect about 10% of the motors to fail by 22,000 hours and 90% of the machines to have failed by 100,000 hours. No further calculations are necessary.

Chapter 570

Life-Table Analysis

Introduction

A life table presents the proportion surviving, the cumulative hazard function, and the hazard rates of a large group of subjects followed over time. The analysis accounts for subjects who die (fail) as well as subjects who are censored (withdrawn). The life-table method competes with the Kaplan-Meier product-limit method as a technique for survival analysis. The life-table method was developed first, but the Kaplan-Meier method has been shown to be superior and with the advent of computers is now the method of choice. However, for large samples, the life-table method is still popular in that it provides a simple summary of a large set of data.

Construction of a Life Table

We will give a brief introduction to the subject in this section. For a complete account of life-table analysis, we suggest the books by Lee (1992) and Elandt-Johnson and Johnson (1980). We will use the same terminology as in the Kaplan-Meier Survival Curves chapter. We suggest that you read the introduction to survival analysis given in that chapter if you are not familiar with common survival analysis terms such as *cumulative survival distribution*, *cumulative hazard function*, and *hazard rates*.

A life table is constructed from a set of grouped or ungrouped failure data. The columns of the table are created using a set of formulas. The rows of the table represent various time intervals. We will now define each of the columns in the life table. Note, however, that because of the large number of columns required to display all of the items, there will be several output reports produced.

Time Interval

Each time interval is represented by $T_t \le T < T_{t+1}$ or $[T_t, T_{t+1})$, where $t = 1, \cdots, s$. The interval is from T_t up to but not including T_{t+1}. The intervals are assumed to be fixed. The intervals do not have to be of equal length, but it is often convenient to make them so.

The midpoint of the interval, T_{mt}, is defined as half way through the interval.

The width of the interval is b_t where $b_t = T_{t+1} - T_t$. The width of the last interval, b_s, is theoretically infinite, so items requiring this value will be left blank.

Number Lost to Follow-Up

The number lost to follow-up, l_t, is the number of individuals who were loss to observation during this interval, so their survival status is unknown.

Number Withdrawn Alive

The number withdrawn alive, w_t, is the number of individuals who had not died (failed) by the end of the study.

Number Dying

The number dying, d_t, is the number of individuals who die (fail) during the interval.

Number Entering the tth Interval

The number entering the *t*th interval, n_t', is computed as follows. In the first interval, it is the total sample size. In the remaining intervals, it is computed using the formula

$$n_t' = n_{t-1}' - l_{t-1} - w_{t-1} - d_{t-1}$$

Number Exposed to Risk

The number exposed to risk, n_t, is computed using the formula

$$n_t = n_{t-1}' - \tfrac{1}{2}\left(l_{t-1} + w_{t-1}\right)$$

This formula assumes that times to loss or withdrawal are distributed uniformly across the interval.

Conditional Proportion Dying

The conditional proportion dying, q_t, is an estimate of the conditional probability of death in the interval given exposure to the risk of death in the interval. It is computed using the formula

$$q_t = \frac{d_t}{n_t}$$

Conditional Proportion Surviving

The conditional proportion surviving, p_t, is an estimate of the conditional probability of surviving through the interval. It is computed using the formula

$$p_t = 1 - q_t$$

Cumulative Proportion Surviving

The cumulative proportion surviving, $S(T_t)$, is an estimate of cumulative survival rate at time T_t. It is computed using the formula

$$S(T_t) = S(T_{t-1})p_{t-1}$$

where

$$S(T_1) = 1$$

The variance of this estimate is itself estimated using the formula

$$V[S(T_t)] = S(T_t)^2 \sum_{j=1}^{t-1} \frac{q_j}{n_j p_j}$$

Using these estimates, pointwise confidence intervals are given using the Kaplan-Meier product-limit formulas given in the Kaplan-Meier chapter.

Estimated Death Density Function

The estimated death density function, $f(T_{mt})$, is an estimate of the probability of dying in this interval per unit width. At the interval midpoint, it is computed using the formula

$$\begin{aligned} f(T_{mt}) &= \frac{S(T_t) - S(T_{t+1})}{b_t} \\ &= \frac{S(T_t) q_t}{b_t} \end{aligned}$$

The variance of this estimate is itself estimated using the formula

$$V[f(T_{mt})] = \frac{S(T_t)^2 q_t^2}{b_t} \sum_{j=1}^{t-1} \left(\frac{q_j}{n_j p_j} + \frac{p_j}{n_j q_j} \right)$$

Hazard Rate Function

The estimated hazard rate function, $h(T_{mt})$, is an estimate of the number of deaths per unit time divided by the average number of survivors at the interval midpoint. It is computed using the formula

$$\begin{aligned} h(T_{mt}) &= \frac{f(T_{mt})}{S(T_{mt})} \\ &= \frac{d_t}{b_t\left(n_t - \frac{1}{2} d_t\right)} \\ &= \frac{2q_t}{b_t(1 + p_t)} \end{aligned}$$

The variance of this estimate is itself estimated using the formula

$$V[h(T_{mt})] = \frac{h(T_{mt})^2}{n_t q_t} \left(1 - \left[\frac{h(T_{mt}) b_t}{2_t} \right]^2 \right)$$

Using these estimates, pointwise confidence intervals are given using the cumulative hazard confidence interval formulas given in the Kaplan-Meier chapter.

Cumulative Hazard Function

The cumulative hazard function, $H(T_t)$, is estimated using the Nelson-Aalen method. It is computed using the formula

$$\tilde{H}(T_t) = \sum_{j=1}^{t} \frac{d_j}{n_j}$$

The variance of this estimate is itself estimated using the formula

$$V\left[\tilde{H}(T_t)\right] = \sum_{j=1}^{t} \frac{d_j}{n_j^2}$$

Using these estimates, pointwise confidence intervals are given using the Nelson-Aalen formulas given in the Kaplan-Meier chapter.

Median Remaining Lifetime

The median remaining lifetime, MRT_t, is the time value at which exactly one-half of those who survived until T_t are still alive.

To compute this value, find the value *j* such that $S(T_j) \geq \frac{1}{2} S(T_t)$ and $S(T_{j+1}) < \frac{1}{2} S(T_t)$. Next, computed the median remaining lifetime using the formula

$$MRT_t = (T_j - T_t) + \frac{b_j\left(S(T_j) - \frac{1}{2}S(T_t)\right)}{S(T_j) - S(T_{j+1})}$$

The variance of this estimate is itself estimated using the formula

$$V(MRT_t) = \frac{S(T_t)^2}{4n_i\left[f(T_{mj})\right]^2}$$

Using these estimates, pointwise confidence intervals are given using the Nelson-Aalen formulas given in the Kaplan-Meier chapter. Note that in this case, the confidence intervals are very crude since the MRT_t are not necessarily distributed normally, even in large samples.

Data Structure

Survival datasets require the ending survival time and an indicator of whether an observation was censored or failed. Additionally, you may also include a frequency variable the gives the count for each row.

The table below shows a dataset from which Lee (1992) constructs a life table. The survival experience of 2418 males with angina pectoris is recorded in years. The life table will use 16 intervals of one year each. These data are contained in the LEE91 database. Note that two rows are required for each data value, one for the failed individuals and another for the censored individuals.

MOTORS dataset

Time	Censor	Count
0.5	1	456
1.5	1	226
2.5	1	152
3.5	1	171
4.5	1	135
5.5	1	125
6.5	1	83
7.5	1	74
8.5	1	51
9.5	1	42
10.5	1	43
11.5	1	34
12.5	1	18
13.5	1	9
14.5	1	6
15.5	1	0
0.5	0	0
1.5	0	39
2.5	0	22
3.5	0	23
4.5	0	24
5.5	0	107
6.5	0	133
7.5	0	102
8.5	0	68
9.5	0	64
10.5	0	45
11.5	0	53
12.5	0	33
13.5	0	27
14.5	0	23
15.5	0	30

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Variables

Time Variable

This variable contains the length of time that an individual was observed. This may represent a failure time or a censor time. Whether the subject actually died is specified by the Censor Variable. Since the values are elapsed times, they must be positive. Zeroes and negative values

are treated as missing values. If you have date values, you much subtract them so that you have a column of elapsed times.

Time Interval Boundaries

Specify a list of times to be used as boundary points along the time scale. These become the left boundaries of the time intervals. Care should be taken to specify a left-most boundary that is less than the smallest time value. This number is often zero.

It is often convenient to make all intervals of the same width, but it is not necessary to do so.

Each interval is closed on the left and open on the right. That is, the interval is T(i) <= T < T(i+1).

Numbers representing the times are separated by blanks or commas. Specify sequences with a colon, putting the increment inside parentheses. For example: 5:25(5) means 5 10 15 20 25. Avoid negative numbers.

Use '(10)' alone to specify ten, equal-spaced values between zero and the maximum.

Censor Variable

Censor Variable

The values in this variable indicate whether the value of the Time Variable represents a censored time or a failure time. These values may be text or numeric. The interpretation of these codes is specified by the Failed and Censored options to the right of this option.

Only two values are used, the Failure code and the Censor code. The Unknown Type option specifies what is to be done with values that do not match either the Failure code or the Censor code.

Rows with missing values (blanks) in this variable are omitted.

Failed

This value identifies those values of the Censor Variable that indicate that the Time Variable gives a failure time. The value may be a number or a letter.

We suggest the letter 'F' or the number '1' when you are in doubt as to what to use.

Censored

This value identifies those values of the Censor Variable that indicate that the individual recorded on this row was censored. That is, the actual failure time occurs sometime after the value of the Time Variable.

We suggest the letter 'C' or the number '0' when you are in doubt as to what to use.

A censored observation is one in which the time until the event of interest is not known because the individual withdrew from the study, the study ended before the individual failed, or for some similar reason.

Unknown Censor

This option specifies what the program is to assume about rows whose censor value is not equal to either the Failed code or the Censored code. Note that observations with missing censor values are always treated as missing.

- **Censored**

  Observations with unknown censor values are assumed to have been censored.

- **Failed**

  Observations with unknown censor values are assumed to have failed.

- **Missing**

  Observations with unknown censor values are assumed to be missing and they are removed from the analysis.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Frequency Variable

Frequency Variable

This variable gives the count, or frequency, of the time displayed on that row. This is the number of subjects represented by each row. When omitted, each row receives a frequency of one. Frequency values should be positive integers.

Options

Confidence Limits

This option specifies the method used to estimate the confidence limits of the survival and hazard values that are displayed. The options are:

- **Linear**

  This is the classical method which uses Greenwood's estimate of the variance.

- **Log Transform**

  This method uses the logarithmic transformation of Greenwood's variance estimate. It produces better limits than the Linear method and has better small sample properties.

- **ArcSine**

  This method uses the arcsine square-root transformation of Greenwood's variance estimate to produce better limits.

Variance

The option specifies which estimator of the variance of the Nelson-Aalen cumulative hazard estimate is to be used. Three estimators have been proposed. When there are no event-time ties, all three give about the same results.

We recommend that you use the Simple estimator unless ties occur naturally in the theoretical event times.

- **Simple**

  This estimator should be used when event-time ties are caused by rounding and lack of measurement precision. This estimate gives the largest value and hence the widest, most conservation, confidence intervals.

- **Plug In**

  This estimator should be used when event-time ties are caused by rounding and lack of measurement precision.

- **Binomial**

  This estimator should be used when ties occur in the theoretical distribution of event times.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary - Median Remaining Lifetime

These options indicate whether to display the corresponding report.

Report Options

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. Single precision is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, only value labels, or both for values of the group variable. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Male, 2=Female, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Decimal Places

Time

This option specifies the number of decimal places shown on reported time values.

Probability

This option specifies the number of decimal places shown on reported probability and hazard values.

N

This option specifies the number of decimal places shown on the number exposed.

Plots Tab

The following options control the plots that are displayed.

Select Plots

Survival/Reliability Plot - Hazard Rate Plot

These options specify which plots type of plots are displayed. Check the plots that you want to see.

Select Plots – Plots Displayed

Individual-Group Plots

When checked, this option specifies that a separate chart of each designated type is displayed.

Combined Plots

When checked, this option specifies that a chart combining all groups is to be displayed.

Plot Options – Plot Arrangement

These options control the size and arrangement of the plots.

Two Plots Per Line

When a man charts are specified, checking this option will cause the size of the charts to be reduced so that they can be displayed two per line. This will reduce the overall size of the output.

Plot Options – Plot Contents

These options control objects that are displayed on all plots.

Function Line

Indicate whether to display the estimated survival (Kaplan-Meier) or hazard function on the plots.

C.L. Lines

Indicate whether to display the confidence limits of the estimated function on the plots.

Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the name of the group variable.

Horizontal (Time) Axis

These options control the horizontal axis of the plots.

Label

This is the text of the horizontal label. The characters *{X}* are replaced the name of the time variable. Press the button on the right of the field to specify the font of the text.

Number of Intervals

This option specifies the number of points along the time axis at which calculations are made. This controls the resolution of the plots. Usually, a value between 50 and 100 is sufficient.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the horizontal (X) axes. If left blank, these values are calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Survival Plot Tab

These options control the attributes of the survival curves. Note that the horizontal axis is specified in the Plots tab.

Vertical Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Survival Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, and *{Z}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Cum Haz Plot Tab

These options control the attributes of the cumulative hazard function plot.

Vertical Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Cum Hazard Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, and *{Z}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Haz Rt Plot Tab

These options control the attributes of the hazard rate plot.

Vertical Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks - Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Hazard Rate Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Title Line 1 and 2

These are the text lines of the titles. The characters *{Y}*, *{X}*, and *{Z}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Lines Tab

These options specify the attributes of the lines used for each group in the hazard curves and survival curves.

Plotting Lines

Line 1 - 15

These options specify the color, width, and pattern of the lines used in the plots of each group. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Labels Tab

The options on this tab specify the labels that a printed on the reports and plots.

Report and Plot Labels

Failure Time Label - Hazard Rate Label

These options specify the term(s) used as labels for these items on the plots. Since these reports are used for performing survival analysis in medical research and reliability analysis in industry, and since these fields often use different terminology, these options are needed to provide appropriate headings for the reports.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**
  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful.

Data Storage Options – Select Items to Store on the Spreadsheet

Group - Median R.L. UCL

Indicate whether to store these values, beginning at the variable indicated by the *Store First Variable In* option.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Life Table

This section presents an example of how to create a life table. This example will use the survival data contained in the LEE91 dataset .

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Life-Table Analysis window.

1 Open the LEE91 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LEE91.S0**.
- Click **Open**.

2 Open the Life-Table Analysis window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Life-Table Analysis**. The Life-Table Analysis procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Life-Table Analysis window, select the **Variables tab**.
- Set the Time Variable to **Time**.
- Set the Time Interval Boundaries to **0:15(1)**.
- Set the Censor Variable to **Censor**.
- Set the Frequency Variable to **Count**.

4 Specify the plots.

- On the Life-Table Analysis window, select the **Plots tab**.
- Check the Hazard Function Plot box.
- Check the **Hazard Rate Plot** box.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum
Failed	15	1625	0.5	14.5
Censored	15	793	1.5	15.5
Total	30	2418	0.5	15.5

This report displays a summary of the amount of data that were analyzed. Scan this report to determine if there were any obvious data errors by double checking the counts and the minimum and maximum times.

Life-Table Analysis Detail Section

Time	No. Start Int'l	No. Lost	No. Died	No. Exp'd to Risk	Conditonal Proportion Surviving	Cumulative Proportion Surviving	Hazard Rate
0.0	2418	0	456	2418.0	0.81141	1.00000	0.20822
1.0	1962	39	226	1942.5	0.88366	0.81141	0.12353
2.0	1697	22	152	1686.0	0.90985	0.71701	0.09441
3.0	1523	23	171	1511.5	0.88687	0.65237	0.11992
4.0	1329	24	135	1317.0	0.89749	0.57856	0.10804
5.0	1170	107	125	1116.5	0.88804	0.51926	0.11860
6.0	938	133	83	871.5	0.90476	0.46112	0.10000
7.0	722	102	74	671.0	0.88972	0.41721	0.11672
8.0	546	68	51	512.0	0.90039	0.37120	0.10483
9.0	427	64	42	395.0	0.89367	0.33422	0.11230
10.0	321	45	43	298.5	0.85595	0.29868	0.15523
11.0	233	53	34	206.5	0.83535	0.25566	0.17942
12.0	146	33	18	129.5	0.86100	0.21356	0.14938
13.0	95	27	9	81.5	0.88957	0.18388	0.11688
14.0	59	23	6	47.5	0.87368	0.16357	0.13483
15.0	30	30	0	15.0	1.00000	0.14291	

This report displays the standard life table. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$.

No. Start Int'l

This is the number entering the *t*th interval. In the first interval, it is the total sample size.

No. Lost

This is the number lost to follow-up and the number withdrawn from the study alive.

No. Died

This is the number of individuals who died (failed) during the interval.

No. Exp'd to Risk

This is the average number exposed to risk in the interval. It is calculated under the assumption that losses and withdrawals are distributed uniformly across the interval.

Conditional Proportion Surviving

This is the conditional proportion surviving through the interval.

Cumulative Proportion Surviving

This is the estimate of the survivorship function, $S(T_t)$. It is also called the cumulative survival rate at time T_t. It is the probability of surviving to the start of the interval and then through the interval.

Hazard Rate

This is the estimated hazard rate function, $h(T_{mt})$. It is an estimate of the number of deaths per unit time divided by the average number of survivors computed at the interval midpoint.

Life-Table Analysis Summary Section

Time	Cumulative Proportion Surviving	Cumulative Hazard Function	Hazard Rate	Death Density Function	Median Remaining Lifetime	No. Starting Int'l
0.0	1.00000	0.18859	0.20822	0.18859	5.3	2418
1.0	0.81141	0.30493	0.12353	0.09440	6.2	1962
2.0	0.71701	0.39508	0.09441	0.06464	6.3	1697
3.0	0.65237	0.50822	0.11992	0.07380	6.2	1523
4.0	0.57856	0.61072	0.10804	0.05931	6.2	1329
5.0	0.51926	0.72268	0.11860	0.05813	5.9	1170
6.0	0.46112	0.81792	0.10000	0.04392	5.6	938
7.0	0.41721	0.92820	0.11672	0.04601	5.2	722
8.0	0.37120	1.02781	0.10483	0.03697	4.9	546
9.0	0.33422	1.13414	0.11230	0.03554	4.8	427
10.0	0.29868	1.27819	0.15523	0.04303	4.7	321
11.0	0.25566	1.44284	0.17942	0.04209		233
12.0	0.21356	1.58184	0.14938	0.02968		146
13.0	0.18388	1.69227	0.11688	0.02031		95
14.0	0.16357	1.81858	0.13483	0.02066		59
15.0	0.14291	1.81858				30

This report displays the most interesting quantities from a life table. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$.

Cumulative Survival

This is the estimate of the survivorship function, $S(T_t)$. It is also called the cumulative survival rate at time T_t. It is the probability of surviving to the start of the interval and then through the interval.

Cumulative Hazard Function

This is the estimate of the cumulative hazard function, $H(T_t)$.

Hazard Rate

This is the estimated hazard rate function, $h(T_{mt})$. It is an estimate of the number of deaths per unit time divided by the average number of survivors computed at the interval midpoint.

Death Density Function

This is the estimated death density function, $f(T_{mt})$. It is an estimate of the probability of dying at the interval midpoint.

Median Remaining Lifetime

This is the median remaining lifetime, MRT_t. It is the time value at which exactly one-half of those who survived until T_t are still alive.

No. Start Int'l

This is the number entering the *t*th interval. In the first interval, it is the total sample size.

Survival Analysis Section

Time	Cumulative Survival	Standard Error	Lower 95% C.L.	Upper 95% C.L.
0.0	1.00000	0.00000	1.00000	1.00000
1.0	0.81141	0.00796	0.79582	0.82701
2.0	0.71701	0.00918	0.69902	0.73500
3.0	0.65237	0.00973	0.63329	0.67145
4.0	0.57856	0.01014	0.55869	0.59844
5.0	0.51926	0.01030	0.49906	0.53945
6.0	0.46112	0.01038	0.44078	0.48147
7.0	0.41721	0.01045	0.39672	0.43769
8.0	0.37120	0.01058	0.35046	0.39193
9.0	0.33422	0.01072	0.31322	0.35523
10.0	0.29868	0.01089	0.27734	0.32003
11.0	0.25566	0.01112	0.23385	0.27746
12.0	0.21356	0.01140	0.19123	0.23590
13.0	0.18388	0.01177	0.16082	0.20694
14.0	0.16357	0.01226	0.13954	0.18760
15.0	0.14291	0.01330	0.11684	0.16898

This report displays the life-table survival distribution along with confidence limits. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$.

Cumulative Survival

This is the estimate of the survivorship function, $S(T_t)$. It is also called the cumulative survival rate at time T_t. It is the probability of surviving to the start of the interval and then through the interval.

Standard Error

This is the large-sample estimate of standard error of the survival function. It is a measure of the precision of the survival estimate.

Lower and Upper Confidence Limits

The lower and upper confidence limits provide a pointwise confidence interval for the survival function. These limit are constructed so that the probability that the true survival probability lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire survival function lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

Cumulative Hazard Section

Time	Cumulative Hazard	Standard Error	Lower 95% C.L.	Upper 95% C.L.
0.0	0.18859	0.00883	0.17128	0.20589
1.0	0.30493	0.01174	0.28192	0.32795
2.0	0.39508	0.01383	0.36797	0.42220
3.0	0.50822	0.01632	0.47624	0.54020
4.0	0.61072	0.01855	0.57437	0.64708
5.0	0.72268	0.02108	0.68137	0.76399
6.0	0.81792	0.02353	0.77180	0.86403
7.0	0.92820	0.02679	0.87568	0.98072
8.0	1.02781	0.03021	0.96860	1.08702
9.0	1.13414	0.03438	1.06676	1.20151
10.0	1.27819	0.04080	1.19824	1.35815
11.0	1.44284	0.04961	1.34560	1.54009
12.0	1.58184	0.05946	1.46531	1.69837
13.0	1.69227	0.06993	1.55521	1.82932
14.0	1.81858	0.08689	1.64829	1.98888
15.0	1.81858	0.08689	1.64829	1.98888

This report displays estimates of the cumulative hazard function. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$.

Cumulative Hazard

This is the Nelson-Aalen estimate of the cumulative hazard function, $H(T_t)$.

Standard Error

This is the estimated standard error of the above cumulative hazard function. The formula used was specified under the Variables tab in the Variance box. The standard error is the square root of this variance.

Lower and Upper Confidence Limits

The lower and upper confidence limits provide a pointwise confidence interval for the cumulative hazard at each time point. These limits are constructed so that the probability that the true cumulative hazard lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire cumulative hazard function lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

Hazard Rate Section

Time	Hazard Rate	Standard Error	Lower 95% C.L.	Upper 95% C.L.
0.0	0.20822	0.00970	0.18921	0.22723
1.0	0.12353	0.00820	0.10746	0.13961
2.0	0.09441	0.00765	0.07942	0.10940
3.0	0.11992	0.00915	0.10197	0.13786
4.0	0.10804	0.00929	0.08984	0.12624
5.0	0.11860	0.01059	0.09784	0.13935
6.0	0.10000	0.01096	0.07851	0.12149
7.0	0.11672	0.01355	0.09017	0.14327
8.0	0.10483	0.01466	0.07610	0.13356
9.0	0.11230	0.01730	0.07839	0.14621
10.0	0.15523	0.02360	0.10898	0.20149
11.0	0.17942	0.03065	0.11935	0.23948
12.0	0.14938	0.03511	0.08056	0.21819
13.0	0.11688	0.03889	0.04065	0.19311
14.0	0.13483	0.05492	0.02719	0.24247
15.0				

This report displays estimates of the hazard rates at the midpoints of each of the time intervals. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$. Note that the hazard rate is actually computed at the midpoint of each interval.

Cumulative Hazard

This is the estimate of the hazard rate, $h(T_{mt})$.

Standard Error

This is the estimated standard error of the above hazard rate. The formula used was given earlier. The standard error is the square root of this variance.

Lower and Upper Confidence Limits

The lower and upper confidence limits provide a pointwise confidence interval for the hazard rate at the midpoint of the time interval. These limits are constructed so that the probability that the true hazard rate lies between them is $1-\alpha$. Note that these limits are constructed for a single time point. Several of them cannot be used together to form a confidence band such that the entire hazard rate lies within the band.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

Death Density Function Section

Time	Death Density	Standard Error	Lower 95% C.L.	Upper 95% C.L.
0.0	0.18859	0.00796	0.17299	0.20418
1.0	0.09440	0.00598	0.08269	0.10612
2.0	0.06464	0.00507	0.05471	0.07458
3.0	0.07380	0.00543	0.06317	0.08444
4.0	0.05931	0.00495	0.04961	0.06900
5.0	0.05813	0.00503	0.04827	0.06800
6.0	0.04392	0.00469	0.03472	0.05311
7.0	0.04601	0.00518	0.03587	0.05615
8.0	0.03697	0.00502	0.02713	0.04682
9.0	0.03554	0.00531	0.02513	0.04594
10.0	0.04303	0.00627	0.03074	0.05532
11.0	0.04209	0.00685	0.02867	0.05551
12.0	0.02968	0.00668	0.01659	0.04278
13.0	0.02031	0.00651	0.00754	0.03307
14.0	0.02066	0.00804	0.00491	0.03641
15.0				

This report displays estimates of the hazard rates at the midpoints of each of the time intervals. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$. Note that the hazard rate is actually computed at the midpoint of each interval.

Death Density

This is the estimate of the death density, $f\left(T_{mt}\right)$.

Standard Error

This is the estimated standard error of the above density. The formula used was given earlier. The standard error is the square root of this variance.

Lower and Upper Confidence Limits

The lower and upper confidence limits provide a pointwise confidence interval for the death density at the midpoint of the time interval. These limits are constructed so that the probability that the true density lies between them is $1-\alpha$.

Three difference confidence intervals are available. All three confidence intervals perform about the same in large samples. The linear (Greenwood) interval is the most commonly used. However, the log-transformed and the arcsine-square intervals behave better in small to moderate samples, so they are recommended. The formulas for these limits were given at the beginning of the chapter and are not repeated here.

Median Remaining Lifetime Section

Time	Median Remaining Lifetime	Standard Error	Lower 95% C.L.	Upper 95% C.L.
0.0	5.3	0.17491	5.0	5.7
1.0	6.2	0.20006	5.9	6.6
2.0	6.3	0.23614	5.9	6.8
3.0	6.2	0.23609	5.8	6.7
4.0	6.2	0.18526	5.9	6.6
5.0	5.9	0.18059	5.6	6.3
6.0	5.6	0.18554	5.2	6.0
7.0	5.2	0.27129	4.6	5.7
8.0	4.9	0.27632	4.4	5.5
9.0	4.8	0.41408	4.0	5.6
10.0	4.7	0.41835	3.9	5.5
11.0				
12.0				
13.0				
14.0				
15.0				

This report displays estimates of the median remaining lifetime of those who are alive at the beginning of the interval. The formulas used were presented earlier.

Time

This is the left boundary of the time interval reported on this line. The right boundary is the entry on the following line. Each interval is represented by $T_t \le T < T_{t+1}$.

Median Remaining Lifetime

This is the estimate of the median remaining lifetime of an individual who survives to the beginning of this interval.

Standard Error

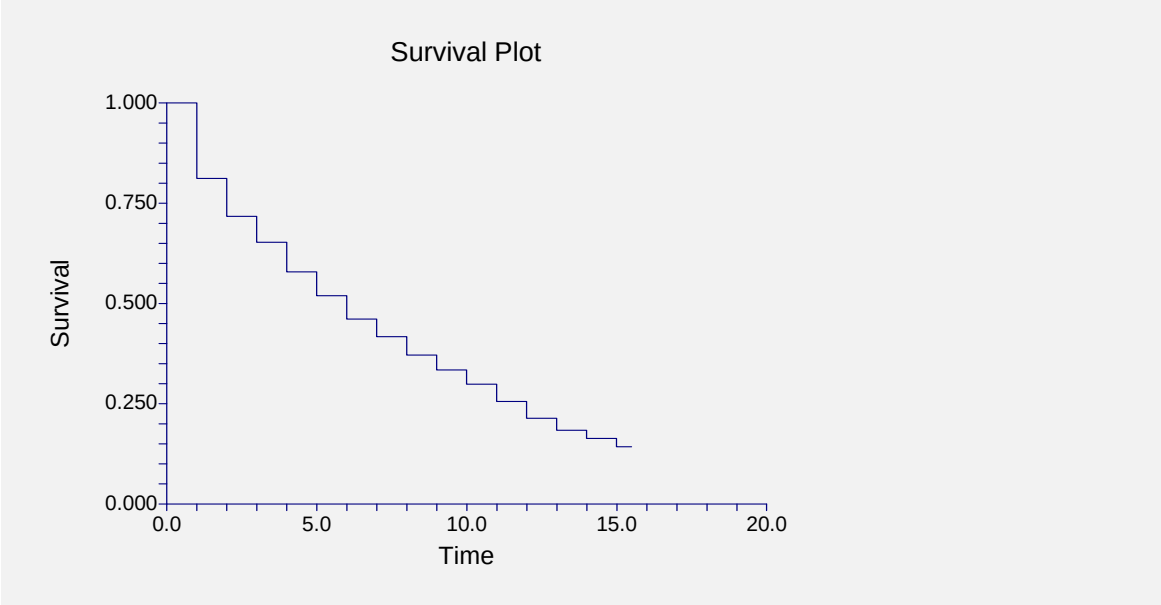
This is the estimated standard error of the above lifetime. The formula used was given earlier.

Lower and Upper Confidence Limits

The lower and upper confidence limits provide a pointwise confidence interval for the hazard rate at the midpoint of the time interval. These limits are constructed so that the probability that the true remaining lifetime lies between them is $1-\alpha$.

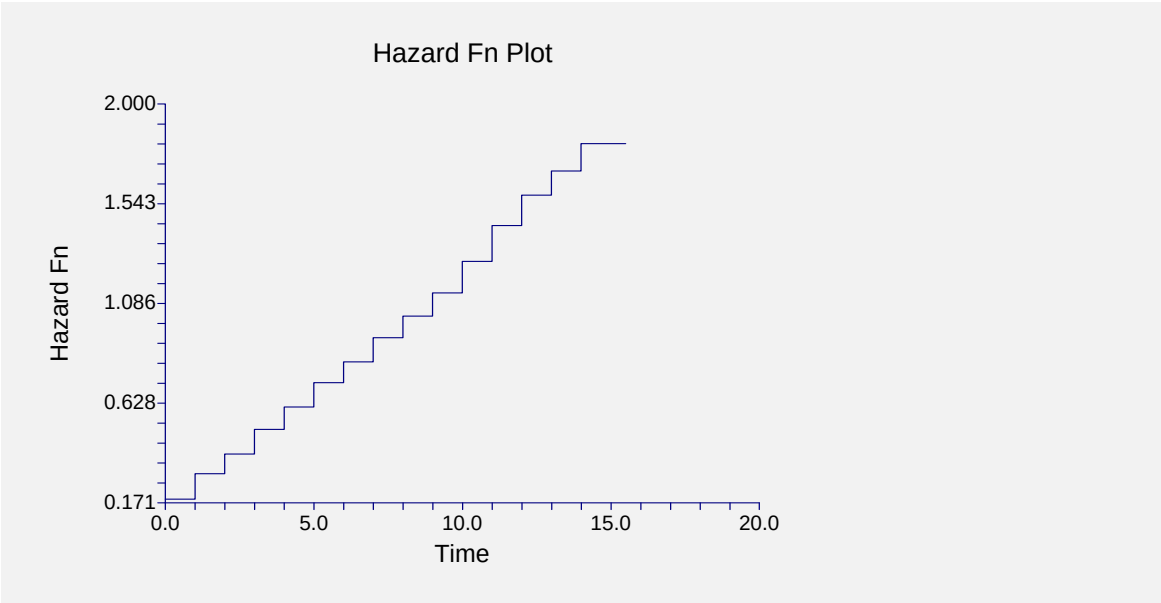
These confidence intervals are highly approximate. They depend on the assumption that the median remaining lifetime is normally distributed. This may not be true even in large samples.

Survival Plot



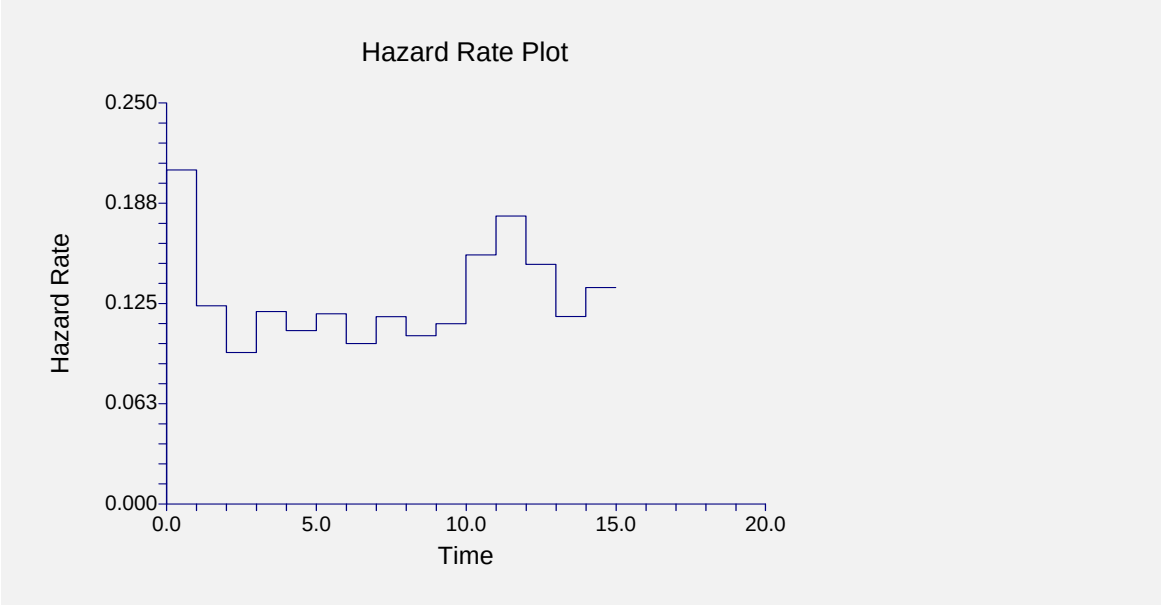
This plot shows the survivorship function. If there are several groups, a separate line is drawn for each group.

Hazard Function Plot



This plot shows the Nelson-Aalen cumulative hazard function. If you have several groups, then a separate line is drawn for each group.

Hazard Rate Plot



This plot shows the hazard rates. Note the unusual step-like appearance of the plot because the hazard rates are assumed constant for the duration of the interval.

Validation of Life-Table Estimator using Lee (1992)

This section presents validation of our life-table estimator. Lee (1992) presents an example on page 91 of a calculated life table. We will include the results of one line of that table so that you can compare those results with those produced by this program. If you compare these values with those shown above, you can validate that ***NCSS*** provides the correct results.

Parameter	Value
Time	3
Lost	23
Dying	171
Entering	1523
Exposed	1511.5
Proportion Dying	0.1131
Proportion Surviving	0.8869
S(T)	0.6524
S.E. S(T)	0.0097
h(T)	0.1199
S.E. h(T)	0.0092
Median Rem. Lifetime	6.23
S.E. MRL(T)	0.9
f(T)	0.0738
S.E. f(T)	0.0054

Chapter 575

Probit Analysis

Introduction

Probit Analysis is a method of analyzing the relationship between a stimulus (dose) and the quantal (all or nothing) response. Quantitative responses are almost always preferred, but in many situations they are not practical. In these cases, it is only possible to determine if a certain response (such as death) has occurred. In a typical quantal response experiment, groups of animals are given different doses of a drug. The percent dying at each dose level is recorded. These data may then be analyzed using Probit Analysis.

The Probit Model assumes that the percent response is related to the log dose as the cumulative normal distribution. That is, the log doses may be used as variables to read the percent dying from the cumulative normal. Using the normal distribution, rather than other probability distributions, influences the predicted response rate at the high and low ends of possible doses, but has little influence near the middle. Hence, much of the comparison of different drugs is done using response rates of fifty percent. The probit model may be expressed mathematically as follows:

$$P = \alpha + \beta\left[\log_{10}\left(Dose\right)\right]$$

where P is five plus the inverse normal transform of the response rate (called the Probit). The five is added to reduce the possibility of negative probits, a situation that caused confusion when solving the problem by hand.

The popularity of the method is due in large part to the work of Finney (1971), in his book Probit Analysis. He explains the proper use and analysis of quantal response data. In **NCSS**, we have coded the algorithms given in his book, and we refer you to it for further information and background.

Data Structure

The data below are suitable for analysis by this procedure. Note that the first variable, Dose, gives the dose level of the treatment. The second variable, Subjects, gives the number of individuals receiving a specific dose level. The third variable, Response, gives the number of treated individuals who exhibited the response of interest.

These data are contained on the SURVIVAL database.

SURVIVAL dataset

Dose	Subjects	Response
50	102	19
60	121	26
70	111	24
80	105	31
90	117	54
100	108	83

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Count Variable

R: Count Variable

This variable contains the number of individuals with the desired response. It must be less than the number of animals. The analysis adds one-half to zero and subtracts one-half if the R = N. This slight modification avoids division by zero in the calculations.

Dose Variable

X: Dose Variable

This option contains the name of the variable containing the dose levels. Note that the analysis uses the log (base 10) transformation of dose levels.

Sample Size Variable

N: Sample Size Variable

This is the variable containing the total number of individuals sampled at a particular dose level.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Reports Tab

The following options control the display of reports and plots.

Select Reports

Percentiles

A separate row in the Dose Percentile report is created for each percentage value given here. This is a list of numbers between 0 and 100 separated by blanks or commas.

Probit Estimation Report ... Dose Percentiles Report

These options specify whether to display the corresponding report.

Select Plots

Dose - Response Plot ... Probit Plot

These options specify whether to display the corresponding plot.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Yes, 2=No, etc.). See the section on specifying Value Labels elsewhere in this manual.

Plot Options

Connect Points on Plots

Checking this option causes all points to be connected with a solid line. This option is useful when you want to study trends across dose levels.

Plot Options – Legend

This section specifies the legend.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the name of the group variable.

Dose - Resp Plot to Probit Plot Tabs

These options control the attributes of the corresponding plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Plot Title

This is the text of the title. The characters {Y} and {X} are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Symbols Tab

These options specify the attributes of the symbols used for each group in the plots.

Plotting Symbols

Group 1 – 15

These options specify the attributes of the symbols used in the plots of each group. The first symbol is used by the first group, the second symbol by the second group, and so on.

Clicking on a symbol box (or the small button to the right of the line box) will bring up a window that allows the attributes to be changed.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Probit Analysis

This section presents an example of how perform a probit analysis using the data that were shown earlier and found in the SURVIVAL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Probit Analysis window.

1 Open the SURVIVAL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SURVIVAL.s0**.
- Click **Open**.

2 Open the Probit Analysis window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Probit Analysis**. The Probit Analysis procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Probit Analysis window, select the **Variables tab**.
- Double-click in the **R: Count Variable** box. This will bring up the variable selection window.
- Select **Response** from the list of variables and then click **Ok**.
- Double-click in the **X: Dose Variable** box. This will bring up the variable selection window.
- Select **Dose** from the list of variables and then click **Ok**.
- Double-click in the **N: Sample Size Variable** box. This will bring up the variable selection window.
- Select **Subjects** from the list of variables and then click **Ok**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Probit Estimation Section

Probit Estimation Section

Parameter	Estimate	Std. Error
Alpha	-4.545974	1.032341
Beta	4.901165	0.5483724
LD50	1.947695	1.304145E-02
Dose50	88.65325	2.662173

Alpha

The estimated value of the intercept, with its associated standard error.

Beta

The estimated value of the slope, with its associated standard error.

LD50

The estimated value, on the log10(dose) scale, at which 50% responded.

Dose50

The estimated value, on the dose scale, at which 50% responded.

Probit Detail Section

Probit Detail Section

Dose	Actual Percent	Probit Percent	N	R	E(R)	Difference	Chi-Square
50	18.63	11.14	102	19.00	11.36	7.64	5.77
60	21.49	20.30	121	26.00	24.56	1.44	0.11
70	21.62	30.75	111	24.00	34.14	-10.14	4.35
80	29.52	41.35	105	31.00	43.41	-12.41	6.05
90	46.15	51.28	117	54.00	60.00	-6.00	1.23
100	76.85	60.12	108	83.00	64.93	18.07	12.62
						Total Chi-Square	30.13
						D.F.	4
						Prob Level	0.00

This report displays a table that would have been used if the calculations were carried out by hand. It is presented more for completeness than for any analytic purpose. It does, however, let you investigate the goodness-of-fit of the dose-response model to the data by considering the Chi-square values.

Dose

The dose level.

Actual Percent

The ratio of the count to the sample size (R/N).

Probit Percent

The estimated ratio (R/N) based on the probit model.

N

The sample size.

R

The count (number responding).

E(R)

The expected count based on the probit model.

Difference

The difference between the actual and the expected counts.

Chi-Square

The Chi-Square statistic for testing the significance (non-zero) of the difference. Since these are single degree of freedom tests, the value should be greater than 3.81 to be significant at the 0.05 level.

Total Chi-Square

The total of the Chi-Square values, used to test the overall significance of the differences from the model.

D.F.

The degrees of freedom of the Chi-Square test.

Prob Level

The probability to the right of the above Chi-Square value. The significance level of the Total Chi-Square test.

Dose Percentile Section

Percentile	Probit	Log(Dose)	Std. Error Log(Dose)	Dose	Std. Error Dose
1	2.6737	1.4730	0.0468	29.7196	3.2008
5	3.3551	1.6121	0.0318	40.9346	2.9993
10	3.7184	1.6862	0.0242	48.5530	2.7013
20	4.1584	1.7760	0.0158	59.7002	2.1685
25	4.3255	1.8101	0.0132	64.5768	1.9640
30	4.4756	1.8407	0.0115	69.2946	1.8364
40	4.7467	1.8960	0.0108	78.7052	1.9529
50	5.0000	1.9477	0.0130	88.6533	2.6622
60	5.2533	1.9994	0.0171	99.8587	3.9219
70	5.5244	2.0547	0.0222	113.4200	5.8064
75	5.6745	2.0853	0.0253	121.7063	7.0888
80	5.8416	2.1194	0.0288	131.6477	8.7309
90	6.2816	2.2092	0.0383	161.8727	14.2814
95	6.6449	2.2833	0.0463	191.9991	20.4873
99	7.3263	2.4223	0.0616	264.4519	37.5022

This report displays the dose levels yielding various predicted response rates.

Percentile

The response rate times 100.

Probit

The normal transform of the percentage plus five. (The five is added to avoid the possibility of a negative probit. This practice was helpful when calculations were done by hand, but is based solely on tradition now that calculations are carried out by computer.)

Log Dose

The logarithm of the dose level (base 10).

Std. Error Log(Dose)

The standard error of the estimated log dose level.

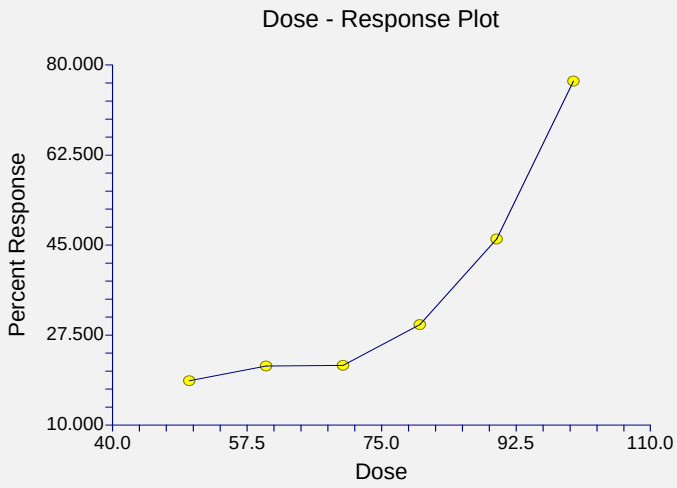
Dose

The dose level.

Std. Error Dose

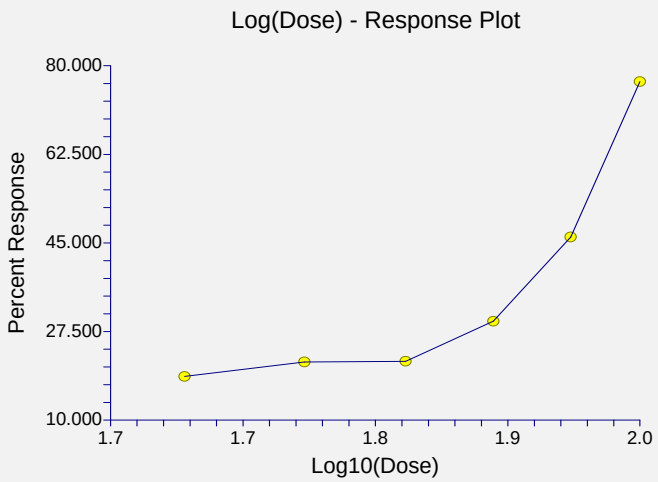
The standard error of the estimated dose level.

Dose-Response Plot



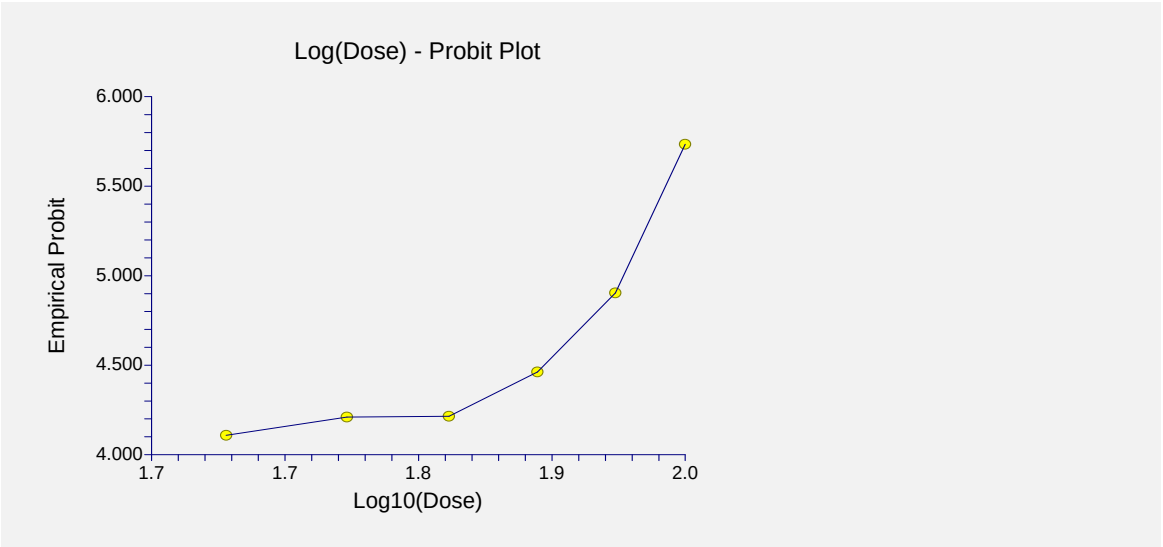
This plot lets you look at the relationship between percent response and dose. Usually, this plot will be nonlinear.

Log(Dose) - Response Plot



This plot lets you look at the relationship between percent response and log dose. Usually, this plot will be nonlinear.

Log(Dose) - Probit Plot



This plot presents the probit model. If the probit model is to be a good approximation, this plot should show a linear relationship. Obviously, in this example, the relationship is quadratic, indicating that the probit model should be modified--perhaps by using the square of Log dose.

Chapter 580

Time Calculator

Introduction

This program module generates and stores elapsed times and censor codes from a database of patient entry, follow-up, and termination dates. It was designed as a supplement for programs like Kaplan-Meier survival analysis and Cox regression which require elapsed times and censor codes as inputs.

Data Structure

This procedure uses up to three date variables to calculate elapsed time and censor codes for each row. These are discussed further below.

Procedure Options

This section describes the options available in this procedure.

Data Tab

Specify the variables to be processed.

Data Variable Specification

Entry (Surgery) Date Variable

This (optional) variable contains the date the subject entered the study. Usually, this corresponds to the surgery or procedure date. The value should be a standard date value such as *mm/dd/yyyy* or *dd/mm/yyyy*. A non-missing value here will override the 'Group Start Date' value.

The elapsed time is calculated by subtracting the entry date from either the Last Follow-Up Date or the Event Date.

Last Follow-Up Date Variable

This variable contains the date the subject was last seen before an event occurred. It is assumed that if, on that visit, the event of interest was seen, the date will be entered in the 'Event Date Variable' and not here. Thus, only subjects who have not exhibited the event should have their times recorded here. The value should be a standard date value such as *mm/dd/yyyy* or *dd/mm/yyyy*. If a 'Group End Date' is specified, this value will override it.

The elapsed time is calculated by subtracting the entry date from either the Last Follow-Up Date or the Event Date. If the Last Follow-Up Date value is non-missing and the Event Date is

missing, the censor variable will be set to a zero (signaling a censored value). Otherwise, the censor variable will be set to a one (signaling an event).

Event (Death) Date Variable

This variable contains the date the subject showed the event (death, remission, etc.). If a subject has not shown the event, this value should be left blank on the database. The value should be a standard date value such as *mm/dd/yyyy* or *dd/mm/yyyy*.

The elapsed time is calculated by subtracting the entry date from either the Last Follow-Up Date or the Event Date. If the Last Follow-Up Date value is non-missing and the Event Date is missing, the censor variable will be set to a zero (signaling a censored value). Otherwise, the censor variable will be set to a one (signaling an event).

Group Date Specification

Group Start Date

If all subjects begin the study on the same date and this date is not on your database, you can enter that date here. The value should be a standard date value such as *mm/dd/yyyy* or *dd/mm/yyyy*. If an 'Entry Date Variable' is specified, its value will override this value.

Group End Date

If follow-up on all subjects ended on the same date and this date is not on your database, you can enter that date here. The value should be a standard date value such as *mm/dd/yyyy* or *dd/mm/yyyy*. If a non-missing 'Last Follow-Up Date Variable' value is entered, it will override this value.

Storage Variable Specification

Elapsed-Time Variable

This variable will receive the calculated elapsed-time. This value can be used as the Time Variable in the Kaplan-Meier or Cox Regression procedures. Note that this value will either be the time to event or the time until end of follow up. The scale of the elapsed time (day, month, or year) is set by the Time Scale option

The elapsed time is calculated by subtracting the entry date from either the Last Follow-Up Date or the Event Date. If the Last Follow-Up Date value is non-missing and the Event Date is missing, the censor variable will be set to a zero (signaling a censored value). Otherwise, the censor variable will be set to a one (signaling an event).

Censor Variable

This variable will receive the censor indicator. A one will appear for all subjects that exhibited the event and a zero will appear for all others. This value can be used as the Censor Variable in the Kaplan-Meier or Cox Regression procedures.

If the Last Follow-Up Date value is non-missing and the Event Date is missing, the censor variable will be set to a zero (signaling a censored value). Otherwise, the censor variable will be set to a one (signaling an event).

Warning: any existing data in this variable will be lost, so choose an empty variable.

Time Scale

Specify the scale that you want to use for the time values.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Preparing Data for Kaplan-Meier Analysis using the Time Calculator

This section presents an example of how to prepare a set of date data for analysis by the Kaplan-Meier procedure. The date values are contained on a database called TIMECALC.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Time Calculator window.

1 Open the TIMECALC dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **TimeCalc.S0**.
- Click **Open**.

2 Open the Time Calculator window.

- On the menus, select **Analysis**, then **Survival / Reliability**, then **Time Calculator**. The Time Calculator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- Select the **Data tab**.
- Set the **Entry Date Variable** to **Entry**.
- Set the **Last Follow-Up Date Variable** to **FollowUp**.
- Set the **Event Date Variable** to **Event**.
- Set the **Elapsed-Time Variable** to **Time**.
- Set the **Censor Variable** to **Censor**.
- Set the **Time Scale** to **Year**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

This procedure does not produce any output. Upon running the procedure, the elapsed times and censor codes will be displayed on the spreadsheet.

Chapter 585

Tolerance Intervals

Introduction

This procedure calculates one-, and two-, sided tolerance intervals based on either a distribution-free (nonparametric) method or a method based on a normality assumption (parametric). A two-sided *tolerance interval* consists of two limits between which a given proportion β of the population falls with a given confidence level $1-\alpha$. A one-sided tolerance interval is similar, but consists of a single upper or lower limit.

Technical Details

Let $X_1, X_2, \cdots, X_n$ be a random sample for a population with distribution function $F(X)$. A $(\beta, 1-\alpha)$ two-sided β-content tolerance interval (T_L, T_U) is defined by

$$\Pr\left[F(T_U) - F(T_L) \geq \beta\right] \geq 1-\alpha$$

A $(\beta, 1-\alpha)$ lower, one-sided β-content tolerance bound T_L is defined by

$$\Pr\left[1 - F(T_L) \geq \beta\right] \geq 1-\alpha$$

A $(\beta, 1-\alpha)$ upper, one-sided β-content tolerance bound T_U is defined by

$$\Pr\left[F(T_U) \geq \beta\right] \geq 1-\alpha$$

Note that a one-sided tolerance limit is the same as the one-sided confidence limit of the quantile of *F*.

Distribution-Free Tolerance Intervals

The definition of two-sided distribution-free tolerance intervals is found in many places. We use the formulation given by Bury (1999). The only distributional assumption made about *F* is that it is a continuous, non-decreasing, probability distribution. That is, these intervals should not be used with discrete data. Given this, the tolerance limits are

$$T_L = X_{(r)}, \quad T_U = X_{(s)}$$

where *r* and *s* are two order indices. The values of *r* and *s* are determined using the formula

$$\sum_{i=0}^{n-2c}\binom{n}{i}\beta^i(1-\beta)^{n-i} \ge 1-\alpha$$

where

$$r = c$$

$$s = n - c + 1$$

The value of c is found as the largest value for which the above inequality is true.

A lower, one-sided tolerance bound is $X_{(r)}$ where r is the largest value for with the following inequality is true.

$$\sum_{i=0}^{n-r}\binom{n}{i}\beta^i(1-\beta)^{n-i} \ge 1-\alpha$$

An upper, one-sided tolerance bound is $X_{(s)}$ where s is the largest value for with the following inequality is true.

$$\sum_{i=0}^{s-1}\binom{n}{i}\beta^i(1-\beta)^{n-i} \ge 1-\alpha$$

Normal-Distribution Tolerance Interval

The limits discussed in this section are based on the assumption that F is the normal distribution.

Two-Sided Limits

In this case, the two-sided tolerance interval is defined by the interval

$$T_L = \bar{x} - ks, \quad T_U = \bar{x} + ks$$

The construction reduces to the determination of the constant k. Howe (1969) provides the following approximation which is 'nearly' exact for all values of n greater than one

$$k = uvw$$

where

$$u = z_{\frac{1+\beta}{2}}\sqrt{1+\frac{1}{n}}$$

$$v = \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}$$

$$w = \sqrt{1+\frac{n-3-\chi^2_{n-1,\alpha}}{2(n+1)^2}}$$

Note that originally, Howe (1969) used $n - 2$ in the above definition of w. But Guenther (1977) gives the corrected version using $n - 3$ shown above.

One-Sided Bound

A one-sided tolerance bound ('bound' is used instead of 'limit' in the one-sided case) is given by

$$T_U = \bar{x} + ks$$

Here k is selected so that

$$\Pr\left(t'_{n-1,\delta} = k\sqrt{n}\right) = 1 - \alpha$$

where $t'_{f,\delta}$ represents a noncentral t distribution with f degrees of freedom and noncentrality

$$\delta = z_\beta \sqrt{n}\,.$$

Data Structure

The data are contained in a single variable.

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Following is a list of the procedure's options.

Variables Tab

The options on this panel specify which variables to use.

Data Variables

Variables

Specify a list of one or more variables for which tolerance intervals are to be generated. You can double-click the field or single click the button on the right of the field to bring up the Variable Selection window.

Group Variable

You can specify a grouping variable. When specified, a separate set of reports is generated for each unique value of this variable.

Exponent

Occasionally, you might want to obtain a statistical report on the square root or log of your variable. This option lets you specify an on-the-fly transformation of the variable. The form of this transformation is $X = Y^A$, where Y is the original value, A is the selected exponent, and X is the resulting value.

Additive Constant

Occasionally, you might want to obtain a statistical report on a transformed version of a variable. This option lets you specify an on-the-fly transformation of the variable. The form of this

transformation is $X = Y+B$, where Y is the original value, B is the specified constant, and X is the value that results.

Note that if you apply both the *Exponent* and the *Additive Constant*, the form of the transformation is $X = (Y + B)^A$.

Frequency Variable

Frequency Variable

This optional variable specifies the number of observations (counts) that each row represents. When omitted, each row represents a single observation. If your data is the result of a previous summarization, you may want certain rows to represent several observations. Note that negative values are treated as a zero count and are omitted.

Population Percentages

Population Percentages for Tolerances

Specify a list of percentages for which tolerance intervals are to be calculated. Note that a tolerance interval is a pair of numbers between which a specified percentage of the population falls. This value is that specified percentage.

In the list, numbers are separated by blanks or commas. Specify sequences with a colon, putting the increment inside parentheses. For example: 5:25(5) means 5 10 15 20 25.

All values in the list must be between 1 and 99.

Data Transformation Options

Exponent

Occasionally, you might want to obtain a statistical report on the square root or log of your variable. This option lets you specify an on-the-fly transformation of the variable. The form of this transformation is $X = Y^A$, where Y is the original value, A is the selected exponent, and X is the resulting value.

Additive Constant

Occasionally, you might want to obtain a statistical report on a transformed version of a variable. This option lets you specify an on-the-fly transformation of the variable. The form of this transformation is $X = Y+B$, where Y is the original value, B is the specified constant, and X is the value that results.

Note that if you apply both the *Exponent* and the *Additive Constant*, the form of the transformation is $X = (Y + B)^A$.

Reports Tab

The options on this panel control the reports and plots displayed.

Select Reports

Descriptive Statistics ... Normality Tests

Indicate whether to display the indicated reports.

Select Plots

Histogram and Probability Plot

Indicate whether to display these plots.

Report Options

Alpha Level

This is the value of alpha for the confidence limits and rejection decisions. Usually, this number will range from 0.1 to 0.001. The default value of 0.05 results in 95% tolerance limits.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. The double-precision option only works when the Decimals option is set to General.

Note that the reports were formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option applies to the *Group Variable*. It lets you select whether to display data values, value labels, or both. Use this option if you want the output to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Decimal Places

Values ... Probabilities Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'General' to display all digits available. The number of digits displayed by this option is controlled by whether the PRECISION option is SINGLE or DOUBLE.

Probability Plot Tab

The options on this panel control the appearance of the probability plot.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* are replaced by the name of the variable. The characters *{M}* are replaced by the name of the selected probability distribution. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on each axis. If left blank, these values are calculated from the data.

Tick Label Settings

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the tick labels along each axis.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Probability Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. The characters *{M}* are replaced by the name of the selected probability distribution. Press the button on the right of the field to specify the font of the text.

Histogram Tab

The options on this panel control the appearance of the histogram.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* are replaced by the name of the variable. The characters *{M}* are replaced by the name of the selected probability distribution. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on each axis. If left blank, these values are calculated from the data.

Tick Label Settings

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the tick labels along each axis.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Histogram Settings

Plot Style File

Designate a histogram style file. This file sets all histogram options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Histogram procedure.

Number of Bars

Specify the number of intervals, bins, or bars used in the histogram.

Titles

Plot Title

This is the text of the title. The characters *{X}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Generating Tolerance Intervals

This section presents a detailed example of how to generate tolerance intervals for the *Height* variable in the SAMPLE data base. To run this example, take the following steps (note that step 1 is not necessary if the SAMPLE dataset is open):

1 Open the SAMPLE dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SAMPLE.S0**.
- Click **Open**.

2 Open the Tolerance Intervals window.

- On the menus, select **Analysis**, then **Descriptive Statistics**, then **Tolerance Intervals**. The Tolerance Intervals procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the options on the Variables tab.

- On the Tolerance Intervals window, select the **Variables tab**. (This is the default.)
- Double-click in the **Variables** text box. This will bring up the variable selection window.
- Select **Height** from the list of variables and then click **Ok**.
- Set the **Population Percentages** to **50 75 80 90 95 99**.

4 Specify the options on the Reports tab.

- On the Tolerance Intervals window, select the **Reports tab**.
- Set the **Decimals-Values** to **3**.
- Set the **Decimals-Means** to **3**.
- Set the **Decimals-Probabilities** to **2**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

The following reports and charts will be displayed in the Output window.

Descriptive Statistics

Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
20	62.100	8.441	1.887	51.000	79.000	28.000

This report was defined and discussed in the Descriptive Statistics procedure chapter. We refer you to the Summary Section of that chapter for details.

Two-Sided Tolerance Intervals

Percent of Population Between Limits	Parametric Lower Tolerance Limit	Parametric Upper Tolerance Limit	Nonparametric Lower Tolerance Limit	Nonparametric Upper Tolerance Limit
50.00	54.074	70.126	52.000	73.000
75.00	48.411	75.789	51.000	79.000
80.00	46.850	77.350		
90.00	42.527	81.673		
95.00	38.777	85.423		
99.00	31.449	92.751		

This section gives the parametric and nonparametric two-sided tolerance intervals.

Percent of Population Between Limits

This is the percentage of population values that are contained in the tolerance interval.

Parametric Lower (Upper) Tolerance Limits

These are the values of the limits of a tolerance interval based on the assumption that the population is normally distributed.

Nonparametric Lower (Upper) Tolerance Limits

These are the values of the limits of a distribution-free tolerance interval. These intervals make no distributional assumption.

Lower One-Sided Tolerance Bounds

Percent of Population Greater Than Bound	Parametric Lower Tolerance Bound	Nonparametric Lower Tolerance Bound
50.00	60.264	56.000
75.00	52.254	52.000
80.00	50.524	51.000
90.00	45.842	
95.00	41.875	
99.00	34.285	

This section gives the parametric and nonparametric one-sided tolerance bounds.

Percent of Population Greater Than Bound

This is the percentage of population values that are above the tolerance bound.

Parametric Lower Tolerance Bound

This is the lower parametric (normal distribution) tolerance bound.

Nonparametric Lower (Upper) Tolerance Limits

This is the lower nonparametric (distribution-free) tolerance bound. Note that some values are missing because of the small sample size in this example.

Upper One-Sided Tolerance Bounds

Percent of Population Less Than Bound	Parametric Upper Tolerance Bound	Nonparametric Upper Tolerance Bound
50.00	63.936	65.000
75.00	71.946	73.000
80.00	73.676	76.000
90.00	78.358	79.000
95.00	82.325	79.000
99.00	89.915	79.000

This section gives the parametric and nonparametric one-sided tolerance bounds.

Percent of Population Less Than Bound

This is the percentage of population values that are below the tolerance bound.

Parametric Lower Tolerance Bound

This is the upper parametric (normal distribution) tolerance bound.

Nonparametric Lower (Upper) Tolerance Limits

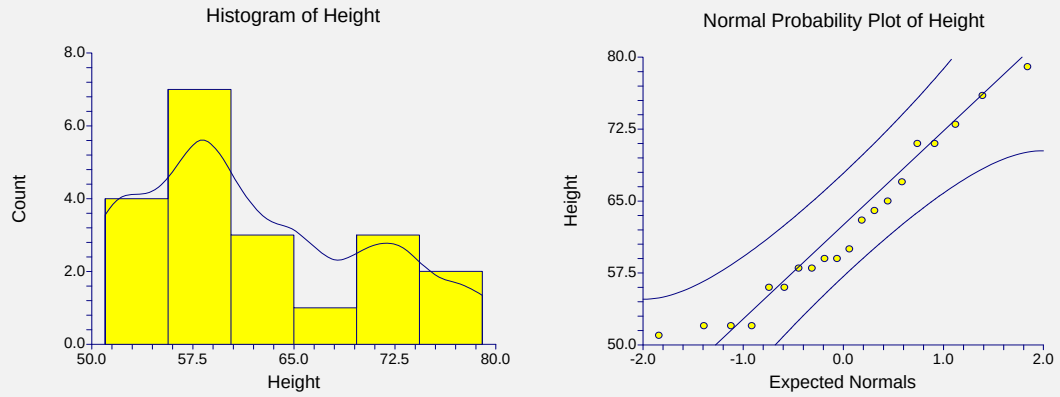
This is the upper nonparametric (distribution-free) tolerance bound.

Normality Test Section

Test Name	Test Value	Prob Level	10% Critical Value	5% Critical Value	Decision (5%)
Shapiro-Wilk W	0.937	0.21			Can't reject normality
Anderson-Darling	0.427	0.31			Can't reject normality
Kolmogorov-Smirnov	0.148		0.176	0.192	Can't reject normality
D'Agostino Skewness	1.037	.30	1.645	1.960	Can't reject normality
D'Agostino Kurtosis	-.7855	.43	1.645	1.960	Can't reject normality
D'Agostino Omnibus	1.6918	.43	4.605	5.991	Can't reject normality

This report was defined and discussed in the Descriptive Statistics procedure chapter. We refer you to the Normality Test Section of that chapter for details.

Plots Section



The plots section displays a histogram and a probability plot to allow you to assess the accuracy of the normality assumption.

References

A

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.

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle," In B. N. Petrov & F. Csaki (Eds.), *The second international symposium on information theory*. Budapest, Hungary: Akademiai Kiado.

Akaike, H. 1974. "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, (6): pages 716-723.

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

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

Al-Sunduqchi, Mahdi S. 1990. *Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics*. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

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

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

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

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

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

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

Armitage, P., and Colton, T. 1998. *Encyclopedia of Biostatistics*. John Wiley, New York.

Armitage,P., McPherson, C.K., and Rowe, B.C. 1969. “Repeated significance tests on accumulating data.” *Journal of the Royal Statistical Society, Series A,* 132, pages 235-244.

Atkinson, A.C. 1985. *Plots, Transformations, and Regression*. Oxford University Press, Oxford (also in New York). This book goes into the details of regression diagnostics and plotting. It puts together much of the recent work in this area.

Atkinson, A.C., and Donev, A.N. 1992. *Optimum Experimental Designs*. Oxford University Press, Oxford. This book discusses D-Optimal designs.

Austin, P.C., Grootendorst, P., and Anderson, G.M. 2007. “A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study,” *Statistics in Medicine*, Volume 26, pages 734-753.

B

Bain, L.J. and Engelhardt, M. 1991. *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker. New York. This book contains details for testing data that follow the exponential and Weibull distributions.

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.

Bates, D. M. and Watts, D. G. 1981. "A relative offset orthogonality convergence criterion for nonlinear least squares," *Technometrics*, Volume 23, 179-183.

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.

Benjamini, Y. and Hochberg, Y. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57, No. 1, 289-300.

Bertsekas, D.P. 1991. *Linear Network Optimization: Algorithms and Codes*. MIT Press. Cambridge, MA.

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.

Bolstad, B.M., et al. 2003. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19, 185-193.

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.

Brown, H., and Prescott, R. 2006. *Applied Mixed Models in Medicine.* 2nd ed. John Wiley & Sons Ltd. Chichester, West Sussex, England.

Brush, Gary G. 1988. *Volume 12: How to Choose the Proper Sample Size*, American Society for Quality Control, 310 West Wisconsin Ave, Milwaukee, Wisconsin, 53203. This is a small workbook for quality control workers.

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).

C

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

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

Casagrande, J. T., Pike, M.C., and Smith, P. G. 1978. "The Power Function of the "Exact" Test for Comparing Two Binomial Distributions," *Applied Statistics*, Volume 27, No. 2, pages 176-180. This article presents the algorithm upon which our Fisher's exact test is based.

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

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

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. 1983. *Graphicals Methods for Data Analysis.* Duxbury Press, Boston, Mass. This wonderful little book is full of examples of ways

to analyze data graphically. It gives complete (and readable) coverage to such topics as scatter plots, probability plots, and box plots. It is strongly recommended.

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

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

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

Chen, T. T. 1997. "Optimal Three-Stage Designs for Phase II Cancer Clinical Trials." *Statistics in Medicine*, Volume 16, pages 2701-2711.

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

Chow, S.C. and Liu, J.P. 1999. *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker. New York.

Chow, S.C.; Shao, J.; Wang, H. 2003. *Sample Size Calculations in Clinical Research*. Marcel Dekker. New York.

Chow, S.-C.; Shao, J.; Wang, H. 2008. *Sample Size Calculations in Clinical Research, Second Edition*. Chapman & Hall/CRC. Boca Raton, Florida.

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.

Cochran, W.G. and Rubin, D.B. 1973. “Controlling bias in observational studies,” *Sankhya, Ser. A*, Volume 35, Pages 417-446.

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.

Cox, D.R. and Snell, E.J. 1981. *Applied Statistics: Principles and Examples*. Chapman & Hall. London, England.

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

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.

D'Agostino, R.B. Jr. 2004. *Tutorials in Biostatistics*. Volume 1. John Wiley & Sons. Chichester, England.

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.

Daniel, W. 1990. *Applied Nonparametric Statistics.* 2nd ed. PWS-KENT Publishing Company. Boston.

Davies, Owen L. 1971. *The Design and Analysis of Industrial Experiments*. Hafner Publishing Company, New York. This was one of the first books on experimental design and analysis. It has many examples and is highly recommended.

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

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

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

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. 1988. “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics*, 44, pages 837-845.

DeMets, D.L. and Lan, K.K.G. 1984. “An overview of sequential methods and their applications in clinical trials.” *Communications in Statistics, Theory and Methods*, 13, pages 2315-2338.

DeMets, D.L. and Lan, K.K.G. 1994. “Interim analysis: The alpha spending function approach.” *Statistics in Medicine*, 13, pages 1341-1352.

Demidenko, E. 2004. *Mixed Models – Theory and Applications*. John Wiley & Sons. Hoboken, New Jersey.

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

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.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. 1994. *Analysis of Longitudinal Data*. Oxford University Press. New York, New York.

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.

Draghici, S. 2003. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC. London. This is an excellent overview of most areas of Microarray analysis.

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.

Dudoit, S., Shaffer, J.P., and Boldrick, J.C. 2003. "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, Volume 18, No. 1, pages 71-103.

Dudoit, S., Yang, Y.H., Callow, M.J., and **Speed, T.P.** 2002. "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Experiments," *Statistica Sinica*, Volume 12, pages 111-139.

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

Dunn, O. J. 1964. "Multiple comparisons using rank sums," *Technometrics*, Volume 6, pages 241-252.

Dunnett, C. W. 1955. "A Multiple comparison procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, Volume 50, pages 1096-1121.

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

Dupont, William. 1988. "Power Calculations for Matched Case-Control Studies," *Biometrics*, Volume 44, pages 1157-1168.

Dupont, William and **Plummer, Walton D.** 1990. "Power and Sample Size Calculations--A Review and Computer Program," *Controlled Clinical Trials*, Volume 11, pages 116-128. Documents a nice public-domain program on sample size and power analysis.

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

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

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

E

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.

F

Farrington, C. P. and Manning, G. 1990. "Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk." *Statistics in Medicine*, Vol. 9, pages 1447-1454. This article contains the formulas used for the Equivalence of Proportions module in PASS.

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

Feldt, L.S.; Ankenmann, R.D. 1999. "Determining Sample Size for a Test of the Equality of Alpha Coefficients When the Number of Part-Tests is Small." *Psychological Methods*, Vol. 4(4), pages 366-377.

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

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

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

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

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, J. L., Levin, B., Paik, M.C. 2003. *Statistical Methods for Rates and Proportions. Third Edition.* John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons. New York. This book provides a very good introduction to clinical trials. It may be a bit out of date now, but it is still very useful.

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

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

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

G

Gans. 1984. "The Search for Significance: Different Tests on the Same Data." *The Journal of Statistical Computation and Simulation*, 1984, pages 1-21.

Gart, John J. and Nam, Jun-mo. 1988. "Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness." *Biometrics*, Volume 44, Issue 2, 323-338.

Gart, John J. and Nam, Jun-mo. 1990. "Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables." *Biometrics*, Volume 46, Issue 3, 637-643.

Gehlback, Stephen. 1988. *Interpreting the Medical Literature: Practical Epidemiology for Clinicians*. Second Edition. McGraw-Hill. New York. Telephone: (800)722-4726. The preface of this book states that its purpose is to provide the reader with a useful approach to interpreting the quantitative results given in medical literature. We reference it specifically because of its discussion of ROC curves.

Gentle, James E. 1998. *Random Number Generation and Monte Carlo Methods*. Springer. New York.

Gibbons, J. 1976. *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart and Winston. New York.

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.

Gu, X.S., and Rosenbaum, P.R. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics*, Vol. 2, No. 4, pages 405-420.

Guenther, William C. 1977. "Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient," *The American Statistician*, Volume 31, Number 1, pages 45-48.

Guenther, William C. 1977. *Sampling Inspection in Statistical Quality Control*. Griffin's Statistical Monographs, Number 37. London.

H

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.

Hochberg, Y. and Tamhane, A. C. 1987. *Multiple Comparison Procedures*. John Wiley & Sons. New York.

Hoerl, A.E. and Kennard, R.W. 1970. "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 55-82.

Hoerl, A.E. and Kennard R.W. 1976. "Ridge regression: Iterative estimation of the biasing parameter." *Communications in Statistics* A5, 77-88.

Howe, W.G. 1969. "Two-Sided Tolerance Limits for Normal Populations—Some Improvements." *Journal of the American Statistical Association*, 64, 610-620.

Hosmer, D. and Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons. New York. This book gives an advanced, in depth look at logistic regression.

Hosmer, D. and Lemeshow, S. 1999. *Applied Survival Analysis*. John Wiley & Sons. New York.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24, 417-441, 498-520.

Hsieh, F.Y. 1989. "Sample Size Tables for Logistic Regression," *Statistics in Medicine*, Volume 8, pages 795-802. This is the article that was the basis for the sample size calculations in logistic regression in PASS 6.0. It has been superceded by the 1998 article.

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression," *Statistics in Medicine*, Volume 17, pages 1623-1634. The sample size calculation for logistic regression in PASS are based on this article.

Hsieh, F.Y. and Lavori, P.W. 2000. "Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates," *Controlled Clinical Trials*, Volume 21, pages 552-560. The sample size calculation for Cox regression in PASS are based on this article.

Hsu, Jason. 1996. *Multiple Comparisons: Theory and Methods*. Chapman & Hall. London. This book gives a beginning to intermediate discussion of multiple comparisons, stressing the interpretation of the various MC tests. It provides details of three popular MC situations: all pairs, versus the best, and versus a control. The power calculations used in the MC module of PASS came from this book.

I

Irizarry, R.A., et al. 2003a. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4, 249-264.

Irizarry, R.A., et al. 2003b. Summaries of Affymetrix GeneChip Probe Level Data. *Nucleic Acids Research*, 31, e15.

J

Jackson, J.E. 1991. *A User's Guide To Principal Components.* John Wiley & Sons. New York. This is a great book to learn about PCA from. It provides several examples and treats everything at a level that is easy to understand.

James, Mike. 1985. *Classification Algorithms*. John Wiley & Sons. New York. This is a great text on the application of discriminant analysis. It includes a simple, easy-to-understand, theoretical development as well as discussions of the application of discriminant analysis.

Jammalamadaka, S.R. and SenGupta, A. 2001. *Topics in Circular Statistics*. World Scientific. River Edge, New Jersey.

Jobson, J.D. 1992. *Applied Multivariate Data Analysis - Volume II: Categorical and Multivariate Methods*. Springer-Verlag. New York. This book is a useful reference for loglinear models and other multivariate methods. It is easy to follows and provides lots of examples.

Jolliffe, I.T. 1972. "Discarding variables in a principal component analysis, I: Artifical data." *Applied Statistics*, 21:160-173.

Johnson, N.L., Kotz, S., and Kemp, A.W. 1992. *Univariate Discrete Distributions, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions Volume 1, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1995. *Continuous Univariate Distributions Volume 2, Second Edition*. John Wiley & Sons. New York.

Jolliffe, I.T. 1986. *Principal Component Analysis*. Springer-Verlag. New York. This book provides an easy-reading introduction to PCA. It goes through several examples.

Julious, Steven A. 2004. "Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data." *Statistics in Medicine*, 23:1921-1986.

Jung, S.-H. 2005. "Sample size for FDR-control in microarray data analysis" *Bioinformatics*, 21(14):3097-3104.

Juran, J.M. 1979. *Quality Control Handbook*. McGraw-Hill. New York.

K

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.

Kenward, M. G. and Roger, J. H. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, pages 983-997.

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.

Kuehl, R.O. 2000. *Design of Experiment: Statistical Principles of Research Design and Analysis, 2nd Edition.* Brooks/Cole. Pacific Grove, California. This is a good graduate level text on experimental design with many examples.

L

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, M.-L. T. 2004. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers. Norwell, Massachusetts.

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.

Little, R. C. et al. 2006. *SAS for Mixed Models – Second Edition*. SAS Institute Inc., Cary, North Carolina.

Liu, H. and Wu, T. 2005. "Sample Size Calculation and Power Analysis of Time-Averaged Difference," *Journal of Modern Applied Statistical Methods*, Vol. 4, No. 2, pages 434-445.

Lu, Y. and Bean, J.A. 1995. "On the sample size for one-sided equivalence of sensitivities based upon McNemar's test," *Statistics in Medicine*, Volume 14, pages 1831-1839.

Lui, J., Hsueh, H., Hsieh, E., and Chen, J.J. 2002. "Tests for equivalence or non-inferiority for paired binary data," *Statistics in Medicine*, Volume 21, pages 231-245.

Lloyd, D.K. and Lipow, M. 1991. *Reliability: Management, Methods, and Mathematics*. ASQC Quality Press. Milwaukee, Wisconsin.

Locke, C.S. 1984. "An exact confidence interval for untransformed data for the ratio of two formulation means," *J. Pharmacokinet. Biopharm.*, Volume 12, pages 649-655.

Lockhart, R. A. & Stephens, M. A. 1985. "Tests of fit for the von Mises distribution." *Biometrika* 72, pages 647-652.

M

Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. *Sample Size Tables for Clinical Studies, 2nd Edition*. Blackwell Science. Malden, Mass. A very good & easy to read book on determining appropriate sample sizes in many situations.

Makridakis, S. and Wheelwright, S.C. 1978. *Iterative Forecasting*. Holden-Day.: San Francisco, California. This is a very good book for the layman since it includes several detailed examples. It is written for a person with a minimum amount of mathematical background.

Manly, B.F.J. 1986. *Multivariate Statistical Methods - A Primer*. Chapman and Hall. New York. This nice little paperback provides a simplified introduction to many multivariate techniques, including MDS.

Mardia, K.V. and Jupp, P.E. 2000. *Directional Statistics*. John Wiley & Sons. New York.

Marple, S.L. 1987. *Digital Spectral Analysis with Applications.* Prentice-Hall: Englewood Cliffs, New Jersey. A technical book about spectral analysis.

Martinez and Iglewicz. 1981. "A test for departure from normality based on a biweight estimator of scale." *Biometrika*, 68, 331-333).

Marubini, E. and **Valsecchi, M.G.** 1996. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley: New York, New York.

Mather, Paul. 1976. *Computational Methods of Multivariate Analysis in Physical Geography.* John Wiley & Sons. This is a great book for getting the details on several multivariate procedures. It was written for non-statisticians. It is especially useful in its presentation of cluster analysis. Unfortunately, it is out-of-print. You will have to look for it in a university library (it is worth the hunt).

Matsumoto, M. and Nishimura,T. 1998. "Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator" *ACM Trans. On Modeling and Computer Simulations*.

Mauchly, J.W. 1940. "Significance test for sphericity of a normal n-variate distribution." *Annals of Mathematical Statistics*, 11: 204-209

McCabe, G.P. 1984. "Principal variables." *Technometrics*, 26, 137-144.

McClish, D.K. 1989. "Analyzing a Portion of the ROC Curve." *Medical Decision Making*, 9: 190-195

McHenry, Claude. 1978. "Multivariate subset selection." *Journal of the Royal Statistical Society, Series C*. Volume 27, No. 23, pages 291-296.

McNeil, D.R. 1977. *Interactive Data Analysis*. John Wiley & Sons. New York.

Mendenhall, W. 1968. *Introduction to Linear Models and the Design and Analysis of Experiments*. Wadsworth. Belmont, Calif.

Metz, C.E. 1978. "Basic principles of ROC analysis." *Seminars in Nuclear Medicine*, Volume 8, No. 4, pages 283-298.

Miettinen, O.S. and Nurminen, M. 1985. "Comparative analysis of two rates." *Statistics in Medicine* 4: 213-226.

Milliken, G.A. and Johnson, D.E. 1984. *Analysis of Messy Data, Volume 1*. Van Nostrand Rienhold. New York, NY.

Milne, P. 1987. *Computer Graphics for Surveying.* E. & F. N. Spon, 29 West 35th St., NY, NY 10001

Montgomery, Douglas. 1984. *Design and Analysis of Experiments*. John Wiley & Sons, New York. A textbook covering a broad range of experimental design methods. The book is not limited to industrial investigations, but gives a much more general overview of experimental design methodology.

Montgomery, Douglas and Peck. 1992. *Introduction to Linear Regression Analysis*. A very good book on this topic.

Montgomery, Douglas C. 1991. *Introduction to Statistical Quality Control.* Second edition. John Wiley & Sons. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 700 pages.

Moore, D. S. and McCabe, G. P. 1999. *Introduction to the Practice of Statistics*. W. H. Freeman and Company. New York.

Mosteller, F. and Tukey, J.W. 1977. *Data Analysis and Regression*. Addison-Wesley. Menlo Park, California. This book should be read by all serious users of regression analysis. Although the terminology is a little different, this book will give you a fresh look at the whole subject.

Motulsky, Harvey. 1995. *Intuitive Biostatistics.* Oxford University Press. New York, New York. This is a wonderful book for those who want to understand the basic concepts of statistical testing. The author presents a very readable coverage of the most popular biostatistics tests. If you have forgotten how to interpret the various statistical tests, get this book!

Moura, Eduardo C. 1991. *How To Determine Sample Size And Estimate Failure Rate in Life Testing*. ASQC Quality Press. Milwaukee, Wisconsin.

Mueller, K. E., and Barton, C. N. 1989. "Approximate Power for Repeated-Measures ANOVA Lacking Sphericity." *Journal of the American Statistical Association,* Volume 84, No. 406, pages 549-555.

Mueller, K. E., LaVange, L.E., Ramey, S.L., and Ramey, C.T. 1992. "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications." *Journal of the American Statistical Association,* Volume 87, No. 420, pages 1209-1226.

Mukerjee, H., Robertson, T., and Wright, F.T. 1987. "Comparison of Several Treatments With a Control Using Multiple Contrasts." *Journal of the American Statistical Association,* Volume 82, No. 399, pages 902-910.

Muller, K. E. and Stewart, P.W. 2006. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley & Sons Inc. Hoboken, New Jersey.

Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, Massachusetts. This is one of the bibles on the topic of regression analysis.

N

Naef, F. et al. 2002. "Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays," *Genome Biol.*, 3, RESEARCH0018.

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

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.

P

Pandit, S.M. and Wu, S.M. 1983. *Time Series and System Analysis with Applications*. John Wiley and Sons. New York. This book provides an alternative to the Box-Jenkins approach for dealing with ARMA models. We used this approach in developing our automatic ARMA module.

Parmar, M.K.B. and Machin, D. 1995. *Survival Analysis*. John Wiley and Sons. New York.

Parmar, M.K.B., Torri, V., and Steart, L. 1998. "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints." *Statistics in Medicine* 17, 2815-2834.

Pearson, K. 1901. "On lines and planes of closest fit to a system of points in space." *Philosophical Magazine* 2, 557-572.

Pan, Z. and Kupper, L. 1999. "Sample Size Determination for Multiple Comparison Studies Treating Confidence Interval Width as Random." *Statistics in Medicine* 18, 1475-1488.

Pedhazur, E.L. and Schmelkin, L.P. 1991. *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. This mammoth book (over 800 pages) covers multivariate analysis, regression analysis, experimental design, analysis of variance, and much more. It provides annotated output from SPSS and SAS which is also useful to our users. The text is emphasizes the social sciences. It provides a "how-to," rather than a theoretical, discussion. Its chapters on factor analysis are especially informative.

Phillips, Kem F. 1990. "Power of the Two One-Sided Tests Procedure in Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 18, No. 2, pages 137-144.

Pocock, S.J. 1977. "Group sequential methods in the design and analysis of clinical trials." *Biometrika*, 64, pages 191-199.

Press, S. J. and Wilson, S. 1978. "Choosing Between Logistic Regression and Discriminant Analysis." *Journal of the American Statistical Association*, Volume 73, Number 364, Pages 699-705. This article details the reasons why logistic regression should be the preferred technique.

Press, William H. 1986. *Numerical Recipes*, Cambridge University Press, New York, New York.

Pregibon, Daryl. 1981. "Logistic Regression Diagnostics." *Annals of Statistics*, Volume 9, Pages 705-725. This article details the extensions of the usual regression diagnostics to the case of logistic regression. These results were extended to multiple-group logistic regression in Lesaffre and Albert (1989).

Price, K., Storn R., and Lampinen, J. 2005. *Differential Evolution – A Practical Approach to Global Optimization.* Springer. Berlin, Germany.

Prihoda, Tom. 1983. "Convenient Power Analysis For Complex Analysis of Variance Models." *Poster Session of the American Statistical Association Joint Statistical Meetings*, August 15-18, 1983, Toronto, Canada. Tom is currently at the University of Texas Health Science Center. This article includes FORTRAN code for performing power analysis.

R

Ramsey, Philip H. 1978 "Power Differences Between Pairwise Multiple Comparisons," *JASA*, vol. 73, no. 363, pages 479-485.

Rao, C.R. , Mitra, S.K., & Matthai, A. 1966. *Formulae and Tables for Statistical Work.* Statistical Publishing Society, Indian Statistical Institute, Calcutta, India.

Ratkowsky, David A. 1989. *Handbook of Nonlinear Regression Models*. Marcel Dekker. New York. A good, but technical, discussion of various nonlinear regression models.

Rawlings John O. 1988. *Applied Regression Analysis: A Research Tool*. Wadsworth. Belmont, California. This is a readable book on regression analysis. It provides a thorough discourse on the subject.

Reboussin, D.M., DeMets, D.L., Kim, K, and Lan, K.K.G. 1992. "Programs for computing group sequential boundaries using the Lan-DeMets Method." Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. John Wiley. New York, New York. This book provides a comprehensive mixture of theoretical and applied results in multivariate analysis. My evaluation may be biased since Al Rencher took me fishing when I was his student.

Robins, Greenland, and Breslow. 1986. "A General Estimator for the Variance of the Mantel-Haenszel Odds Ratio," *American Journal of Epidemiology*, vol.42, pages 719-723.

Robins, Breslow, and Greenland. 1986. "Estimators of the Mantel-Haenszel variance consisten in both sparse data and large-strata limiting models," *Biometrics*, vol. 42, pages 311-323.

Rosenbaum, P.R. 1989. "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, vol. 84, no. 408, pages 1024-1032.

Rosenbaum, P.R., and Rubin, D.B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, pages 41-55.

Rosenbaum, P.R., and Rubin, D.B. 1984. "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, vol. 79, pages 516-524.

Rosenbaum, P.R., and Rubin, D.B. 1985a. "Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician*, vol. 39, pages 33-38.

Rosenbaum, P.R., and Rubin, D.B. 1985b. "The Bias Due to Incomplete Matching," *Biometrics*, vol. 41, pages 106-116.

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.

S

Sahai, Hardeo & Khurshid, Anwer. 1995. *Statistics in Epidemiology*. CRC Press. Boca Raton, Florida.

Schiffman, Reynolds, & Young. 1981. *Introduction to Multidimensional Scaling*. Academic Press. Orlando, Florida. This book goes through several examples.

Schilling, Edward. 1982. *Acceptance Sampling in Quality Control*. Marcel-Dekker. New York.

Schlesselman, Jim. 1981. *Case-Control Studies*. Oxford University Press. New York. This presents a complete overview of case-control studies. It was our primary source for the Mantel-Haenszel test.

Schmee and Hahn. November, 1979. "A Simple Method for Regression Analysis." *Technometrics*, Volume 21, Number 4, pages 417-432.

Schoenfeld, David A. 1983. "Sample-Size Formula for the Proportional-Hazards Regression Model" *Biometrics,* Volume 39, pages 499-503.

Schoenfeld, David A. and **Richter, Jane R.** 1982**.** "Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint," *Biometrics,* March 1982, Volume 38, pages 163-170.

Schork, M. and Williams, G. 1980. "Number of Observations Required for the Comparison of Two Correlated Proportions." *Communications in Statistics-Simula. Computa.,* B9(4), 349-357.

Schuirmann, Donald. 1981**.** "On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval," *Biometrics,* Volume 37, pages 617.

Schuirmann, Donald. 1987**.** "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics,* Volume 15, Number 6, pages 657-680.

Seber, G.A.F. 1984. *Multivariate Observations*. John Wiley & Sons. New York. (This book is an encyclopedia of multivariate techniques. It emphasizes the mathematical details of each technique and provides a complete set of references. It will only be useful to those comfortable with reading mathematical equations based on matrices.)

Seber, G.A.F. and Wild, C.J. 1989. *Nonlinear Regression*. John Wiley & Sons. New York. This book is an encyclopedia of nonlinear regression.

Senn, Stephen. 1993. *Cross-over Trials in Clinical Research*. John Wiley & Sons. New York.

Senn, Stephen. 2002. *Cross-over Trials in Clinical Research.* Second Edition. John Wiley & Sons. New York.

Shapiro, S.S. and Wilk, M.B. 1965 "An analysis of Variance test for normality." *Biometrika*, Volume 52, pages 591-611.

Shuster, Jonathan J. 1990. *CRC Handbook of Sample Size Guidelines for Clinical Trials.* CRC Press, Boca Raton, Florida. This is an expensive book ($300) of tables for running log-rank tests. It is well documented, but at this price it better be.

Signorini, David. 1991. "Sample size for Poisson regression," *Biometrika,* Volume 78, 2, pages 446-450.

Simon, Richard. "Optimal Two-Stage Designs for Phase II Clinical Trials," *Controlled Clinical Trials,* 1989, Volume 10, pages 1-10.

Snedecor, G. and Cochran, Wm. 1972. *Statistical Methods*. The Iowa State University Press. Ames, Iowa.

Sorribas, A., March, J., and Trujillano, J. 2002. "A new parametric method based on S-distributions for computing receiver operating characteristic curves for continuous diagnostic tests." *Statistics in Medicine* 21, 1213-1235.

Spath, H. 1985. *Cluster Dissection and Analysis.* Halsted Press. New York. (This book contains a detailed discussion of clustering techniques for large data sets. It contains some heavy mathematical notation.)

Speed, T.P. (editor). 2003. *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC. Boca Raton, Florida.

Stekel, D. 2003. *Microarray Bioinformatics.* Cambridge University Press. Cambridge, United Kingdom.

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.

T

Tabachnick, B. and Fidell, L. 1989. *Using Multivariate Statistics.* Harper Collins. 10 East 53d Street, NY, NY 10022. This is an extremely useful text on multivariate techniques. It presents computer printouts and discussion from several popular programs. It provides checklists for each procedure as well as sample written reports. I strongly encourage you to obtain this book!

Tango, Toshiro. 1998. "Equivalence Test and Confidence Interval for the Difference in Proportions for the Paired-Sample Design." *Statistics in Medicine*, Volume 17, 891-908.

Therneau, T.M. and Grambsch, P.M. 2000. *Modeling Survival Data*. Springer: New York, New York. A the time of the writing of the Cox regression procedure, this book provides a thorough, up-to-date discussion of this procedure as well as many extensions to it. Recommended, especially to those with at least a masters in statistics.

Thomopoulos, N.T. 1980. *Applied Forecasting Methods*. Prentice-Hall: Englewood Cliffs, New Jersey. This book contains a very good presentation of the classical forecasting methods discussed in chapter two.

Thompson, Simon G. 1998. *Encyclopedia of Biostatistics, Volume 4*. John Wiley & Sons. New York. Article on Meta-Analysis on pages 2570-2579.

Tiku, M. L. 1965. "Laguerre Series Forms of Non-Central X^2 and F Distributions," *Biometrika*, Volume 42, pages 415-427.

Torgenson, W.S. 1952. "Multidimensional scaling. I. Theory and method." *Psychometrika* 17, 401-419. This is one of the first articles on MDS. There have been many advances, but this article presents many insights into the application of the technique. It describes the algorithm on which the metric solution used in this program is based.

Tubert-Bitter, P., Manfredi,R., Lellouch, J., Begaud, B. 2000. "Sample size calculations for risk equivalence testing in pharmacoepidemiology." *Journal of Clinical Epidemiology* 53, 1268-1274.

Tukey, J.W. and McLaughlin, D.H. 1963. "Less Vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization." *Sankhya, Series A* 25, 331-352.

Tukey, J.W. 1977. *Exploratory Data Analysis.* Addison-Wesley Publishing Company. Reading, Mass.

U

Upton, G.J.G. 1982."A Comparison of Alternative Tests for the 2 x 2 Comparative Trial.", *Journal of the Royal Statistical Society*, Series A,, Volume 145, pages 86-105.

Upton, G.J.G. and Fingleton, B. 1989. *Spatial Data Analysis by Example: Categorical and Directional Data. Volume 2.* John Wiley & Sons. New York.

V

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.

W

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.

Welch, B.L. 1947. "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. 1949. "Further Note on Mrs. Aspin's Tables and on Certain Approximations to the Tabled Function," *Biometrika*, 36, 293-296.

Westfall, P. et al. 1999. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc. Cary, North Carolina.

Westgard, J.O. 1981. "A Multi-Rule Shewhart Chart for Quality Control in Clinical Chemistry," *Clinical Chemistry*, Volume 27, No. 3, pages 493-501. (This paper is available online at the www.westgard.com).

Westlake, W.J. 1981. "Bioequivalence testing—a need to rethink," *Biometrics*, Volume 37, pages 591-593.

Whittemore, Alice. 1981. "Sample Size for Logistic Regression with Small Response Probability," *Journal of the American Statistical Association*, Volume 76, pages 27-32.

Wickens, T.D. 1989. *Multiway Contingency Tables Analysis for the Social Sciences.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. A thorough book on the subject. Discusses loglinear models in depth.

Wilson, E.B.. 1927. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, Volume 22, pages 209-212. This article discusses the 'score' method that has become popular when dealing with proportions.

Winer, B.J. 1991**.** *Statistical Principles in Experimental Design (Third Edition)*. McGraw-Hill. New York, NY. A very complete analysis of variance book.

Wit, E., and McClure, J. 2004. *Statistics for Microarrays*. John Wiley & Sons Ltd, Chichester, West Sussex, England.

Wolfinger, R., Tobias, R. and Sall, J. 1994. "Computing Gaussian likelihoods and their derivatives for general linear mixed models," *SIAM Journal of Scientific Computing*, 15, no.6, pages 1294-1310.

Woolson, R.F., Bean, J.A., and Rojas, P.B. 1986. "Sample Size for Case-Control Studies Using Cochran's Statistic," *Biometrics*, Volume 42, pages 927-932.

Y

Yuen, K.K. and Dixon, W. J. 1973. "The approximate behavior and performance of the two-sample trimmed t," *Biometrika*, Volume 60, pages 369-374.

Yuen, K.K. 1974. "The two-sample trimmed t for unequal population variances," *Biometrika*, Volume 61, pages 165-170.

Z

Zar, Jerrold H. 1984**.** *Biostatistical Analysis (Second Edition)*. Prentice-Hall. Englewood Cliffs, New Jersey. This introductory book presents a nice blend of theory, methods, and examples for a long list of topics of interest in biostatistical work.

Zhou, X., Obuchowski, N., McClish, D. 2002**.** *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc. New York, New York. This is a great book on the designing and analyzing diagnostic tests. It is especially useful for its presentation of ROC curves.

Chapter Index

3

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Index

2

3

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z