Statistics and Measurement Concepts for LazStats

A User's Manual

By

William G. Miller

2012

Table of Contents

Table of Figures

Chapter 1. Installing and Using LazStats

Introduction

LazStats, among others, are ongoing projects that I have created for use by students, teachers, researchers, practitioners and others. There is no charge for use of these programs if downloaded directly from a World Wide Web site. The software is a result of an "over-active" hobby of a retired professor (Iowa State University.) I make no claim or warranty as to the accuracy, completeness, reliability or other characteristics desirable in commercial packages (as if they can meet these requirement also.) They are designed to provide a means for analysis by individuals with very limited financial resources. The typical user is a student in a required social science or education course in beginning or intermediate statistics, measurement, psychology, etc. Some users may be individuals in developing nations that have very limited resources for purchase of commercial products.

LazStats was written using the Free Pascal/Lazarus compiler which may be downloaded from their site on the Internet. Versions are available for multiple operating systems such as Windows, Linux, Mac OSX, etc. A program written for one platform can, theoretically, be compiled again for another platform. The package is similar to the previously available Borland Delphi compiler.

While I reserve the copyright protection of these packages, I make no restriction on their distribution or use. It is common courtesy, of course, to give me credit if you use these resources. Because I do not warrant them in any manner, you should insure yourself that the routines you use are adequate for your purposes. I strongly suggest analyses of textbook examples and comparisons to other statistical packages where available. You should also be aware that I am constantly revising, correcting and updating LazStats. For that reason, some of the images and descriptions in this book may not be exactly as you see when you execute the program. I update this book from time to time to try and keep the program and text coordinated.

Installing LazStats

LazStats has been successfully installed on Windows 95, 98, ME, XT, NT, Vista and Windows 7 systems. Other versions have been compiled by Dr. Chris Rorden and may be acquired by visiting my web site at http://www.statprograms4U.com. A free setup package (INNO) has been used to package LazStats for installation on your computer. Included in the setup file are the executable file and HTML files used to access help for various procedures. At this time, only the LazStats version is receiving my attention for updates and revisions. Individuals with other platforms that know some programming are encouraged to download Lazarus and my LazStats source code and build a version of LazStats for their own use.

To install LazStats for Windows, follow these steps:

1. Connect to the internet address: http://www.statprograms4U.com

2. Click on the link to the LazStats INNO setup file. It is about 10 megabytes in size. Your browser should automatically begin the download process to a directory on your computer.

3. Once you have downloaded the INNO setup file, simply double click the name of the file and the setup will begin.

4. By default, Windows will normally install LazStats in a programs directory on the C drive. Several users have had success in using the INNO setup to place the program on a "memory stick" type of drive that plugs into the "USB" port. You are encouraged to select a directory in your main user area such as the "Documents" directory. Simply follow the directions provided by the setup program and complete the installation. When completed, there should be an entry in the Programs menu.

Starting LazStats

To begin using a Windows version of LazStats simply click the Windows “Start” button in the lower left portion of your screen, move the cursor to the “Programs” menu and click on the LazStats entry. The initial screen you see will be a form that displays the “Open Source” license for the use of this package. Please read it! Once you have read it and clicked the button to continue, you will be notified that an “Options” file was created. The OPTIONS.FIL contains the default values for how your data will be stored and what the default missing value is for a variable (more on this later.) Next, the following form should appear:

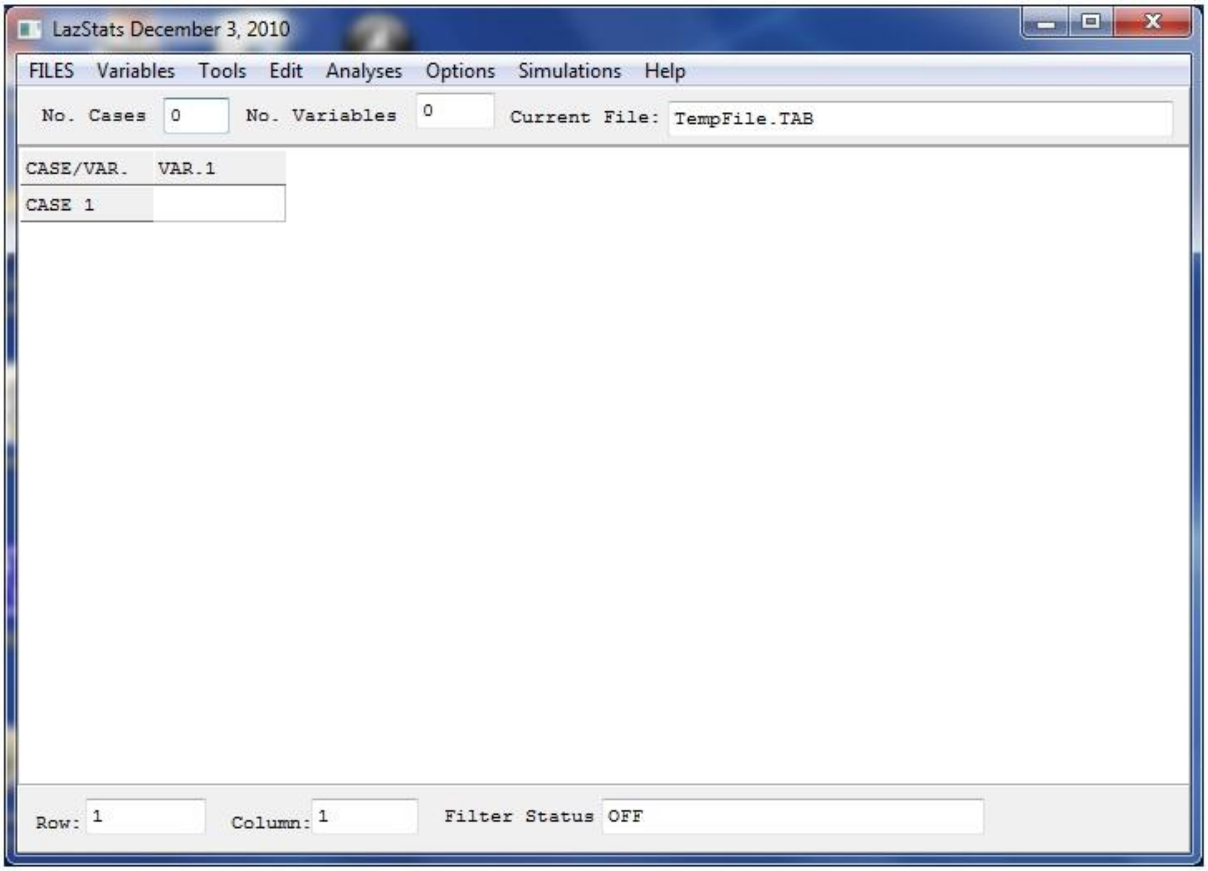


Fig. 1.1 The LazStats Main Form

The above form contains several important areas. The “grid” is where data values are entered. Each column represents a “variable” and each row represents an “observation” or case. A default label is given for the first variable and each case of data you enter will have a case number. At the top of this “main” form there is a series of “drop-down” menu items. When you click on one of these, a series of options (and sometimes sub-options) that you can click to select are shown. Before you begin to enter case values, you probably should “define” each variable to be entered in the data grid. Select the “VARIABLES” menu item and click the “Define” option. More will be said about this in the following pages.

Files

The “heart” of LazStats or any other statistics package is the data file to be created, saved, retrieved and analyzed. Unfortunately, there is no one “best” way to store data and each data analysis package has its own method for storing data. Many packages do, however, provide options for importing and exporting files in a variety of formats. For example, with Microsoft’s Excel package, you can save a file as a file of “tab” separated fields. Other program packages such as SPSS can import “tab” files. Here are the types of file formats supported by LazStats:

1.Text files (with the extension .LAZ) NOTE: the file format in this text file is unique to LazStats!
2.Tab separated field files (with the file extension of .TAB.)
3.Comma separated field files (with the file extension of .CSV.)
4.Space separated field files (with the file extension of .SSV.)

My preference is to save files as either a .TEX or .TAB file. This gives me the opportunity to analyze the same data using a variety of packages. For relatively small files (say, for example, a file with 20 variables and 1000 cases), the speed of loading the different formats is similar and quite adequate.

Creating a File

When LazStats begins, you will see a “grid” of two rows and two columns. The left-most column will automatically contain the word “Case” followed by a number (1 for the first case.) The top row will contain the default name of the first variable. You can change the name of the first variable and define additional variables by clicking on the menu item labeled “VARIABLES” and then clicking on the “Define” option. A “form” will appear that looks like the Fig. below:

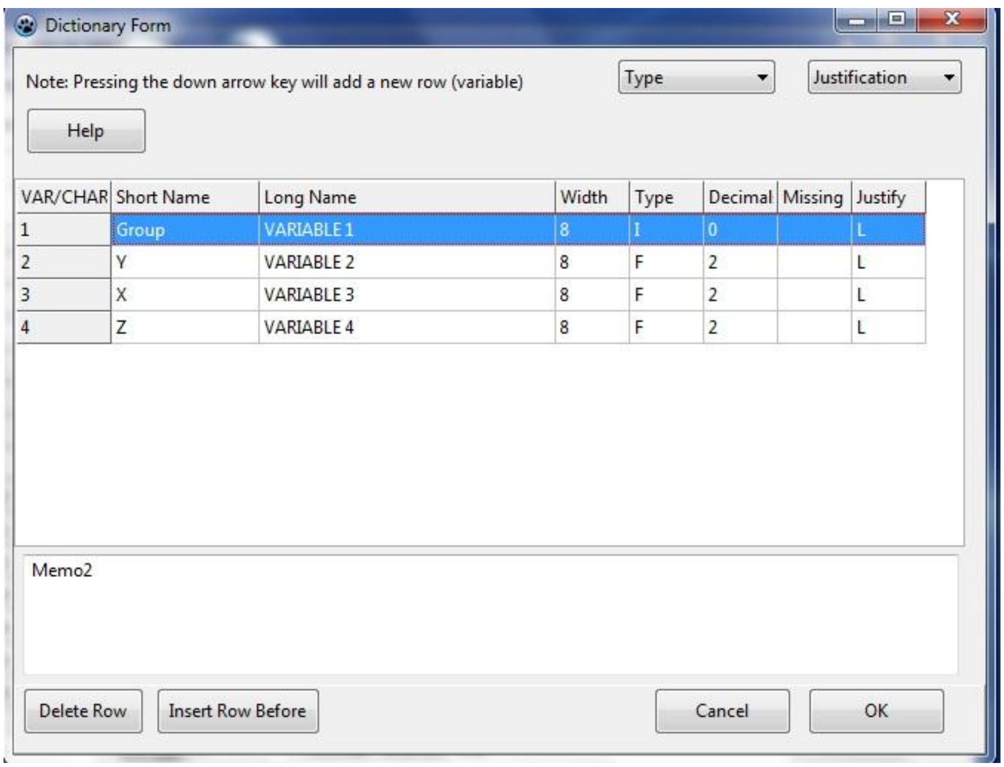


Fig. 1.2 Variable Definitions Dialog

In the above Fig. you will notice that a variable name was automatically generated for the first variable. To change the default name, click the box with the default Short Name and enter the variable name that you desire. It is suggested that you keep the length of the name to eight characters or less. You may also enter a long label for the variable. If you save your file as a .LAZ file, this long name (as well as other descriptive information) will be saved in the file (the use of the long label has not yet been implemented for printing output but will be in future versions.) To proceed, simply click the Return button in the lower right of this form. The default type of variable is a “floating point” value, that is, a number which may contain a decimal fraction. If a data field (grid cell) is left blank, the program will usually assume a missing value for the data. The default format of a data value is eight positions with three positions allocated to fractional decimal values (format 8.3.) By clicking on any of the specification fields you can modify these defaults to your own preferences. You can change the number of decimal places (0 for integers.) You will find that some analyses require that a variable be defined as an integer and others as floating point values. The drop-down box labeled “Var. Types” lets you click on the type of variable you are defining and automatically record the integer value that defines that type. If you press the “down-arrow” on your keyboard, another variable with default values will be added. You can also insert or delete a new variable by clicking one of the buttons at the bottom of the form. Another way to specify the default format and missing values is by modifying the "Options" file. When you click on the Options menu the following form appears:

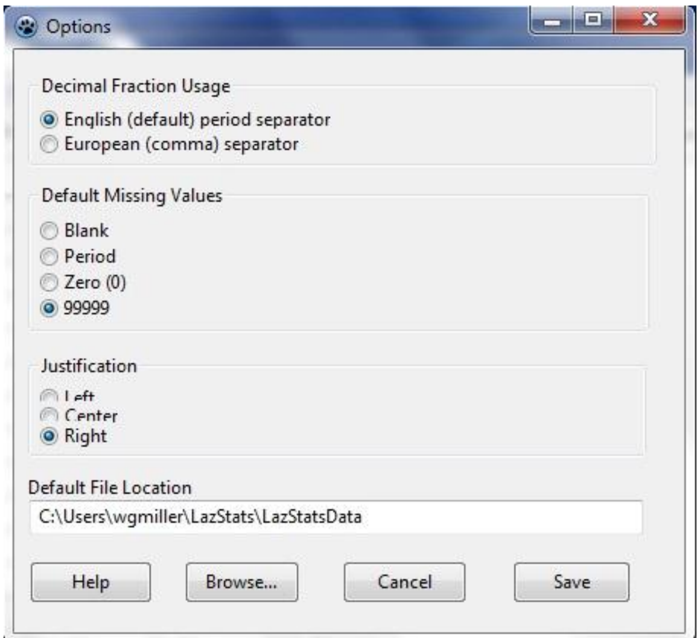


Fig. 1.3 The Options Menu

In the options form you can specify the Data Entry Defaults as well as whether you will be using American or European formatting of your data (American's use a period (.) and Europeans use a comma (,) to separate the integer portion of a number from its fractional part.) To change the path to your data, double click your mouse button on the “Browse” button and then sub-directories of your choice. Double click on that directory to obtain a list of the files in that directory. In many countries, the separation of the whole number from the fractional part of a floating point number is a comma (,) and not a period (.) as in the United States. A user that uses the comma separator is designated a “European” user. The default is the American usage. It is possible to convert one type to another. The example files all use the American standard. If you use the European standard, you will need to examine the “default” confidence intervals shown on many of the statistics dialog forms – they may have a period (e.g. 0.05) instead of a comma (0,05) as needed in the European format. One can click on the value and change it to an appropriate format.

Entering Data

When you enter data in the grid of the main form there are several ways to navigate from cell to cell. You can, of course, simply click on the cell where you wish to enter data and type the data values. If you press the “enter” key following the typing of a value, the program will automatically move you to the next cell to the right of the current one (assuming you have defined more than one variable.) You may also press the keyboard “down” arrow to move to the cell below the current one. If it is a new row for the grid, a new row will automatically be added and the “Case” label added to the first column. You may use the arrow keys to navigate left, right, up and down. You may also press the “Page Up” button to move up a screen at a time, the “Home” button to move to the beginning of a row, etc. Try the various keys to learn how they behave. You may click on the main form’s Edit menu and use the delete column or delete row options. Be sure the cursor is sitting in a cell of the row or column you wish to delete when you use this method. A common problem for the beginner is pressing the "enter" key when in the last column of their variables. If you do accidentally add a case or variable you do not wish to have in your file, use the edit menu and delete the unused row or variable. Notice that as you make grid entries and move to another cell, the previous value is automatically formatted according to the definition for that variable.

Saving a File

Once you have entered a number of values in the grid, it is a good idea to save your work (power outages do occur!) Go to the main form’s File menu and click it. You will see there are several ways to save your data. A “dialog box” will then appear as shown below for a .TEX type of file:

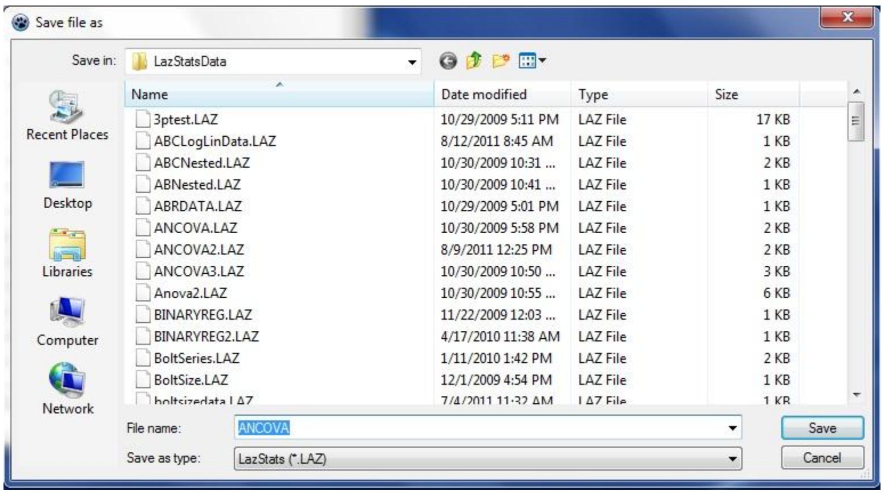


Fig. 1.4 The Save Dialog Form

Simply type the name of the file you wish to create in the File name box and click the Save button. After this initial save operation, you may continue to enter data and save again. Before you exit the program, be sure to save your file if you have made additions to it. If you try to exit the program when a file is still in the grid, you will be asked if you want to save the file before exiting. You can avoid this by closing the file before exiting the program. If you have imported a .TAB file and are now saving it a a .LAZ file, be sure to type the extension .LAZ after the file name.

If you do not need to save specifications other than the short name of each variable, you may prefer to “export” the file in a format compatible to other programs. The Export Tab File option under the File menu will save your data in a text file in which the cell values in each row are separated by a tab key character. A file with the extension .TAB will be created. The list of variables from the first row of the grid are saved first, then the first row of the data, etc. until all grid rows have been saved. If there are blanks in any value cells, the default missing value will be written for that cell. Alternatively, you may export your data with a comma or a space separating the cell values. Basic language programs frequently read files in which values are separated by commas or spaces. If you are using the European format of fractional numbers, DO NOT USE the comma separated files format since commas will appear both for the fractions and the separation of values - clearly a design for disaster!

The Main Form Menus

Help

Users of Microsoft Windows are used to having a “help” system available to them for instant assistance when using a program. Most of these systems provide the user the ability to press the “F1" key for assistance on a particular topic or by placing their cursor on a particular program item and pressing the right mouse button to get help.

LazStats for the Microsoft Windows does not use the MicroSoft help file system. Instead, it uses your Internet browser to display a “Portable Data File” (.PDF) file. Place the cursor on a menu topic and press the F1 key to see what happens! You can use the help system to learn more about LazStats procedures. Again, as the program is revised, there may not yet be help topics for all procedures and some help topics may differ from the actual procedure's operation.

The Variables Menu

Across the top of the "Main Form" is a series of "menu" items. Like the "File" menu, each of these menu items "drops-down" a series of options and these options may have sub-options. The "Variables" menu contains a variety of options to assist you in working with the variables (columns of data) that you enter in the grid. These options include:

1. Define
2. Print Definitions
3. Recode a variable’s values
4. Transform a variable
5. Enter an Equation to Combine Variables to Create a New Variable

The first option lets you enter or change a variable definition (see Fig. 2 above.)

The fourth option lets you "transform" an existing variable to create a new variable. A variety of transformations are possible. If you elect this option, you will see the following dialogue form:

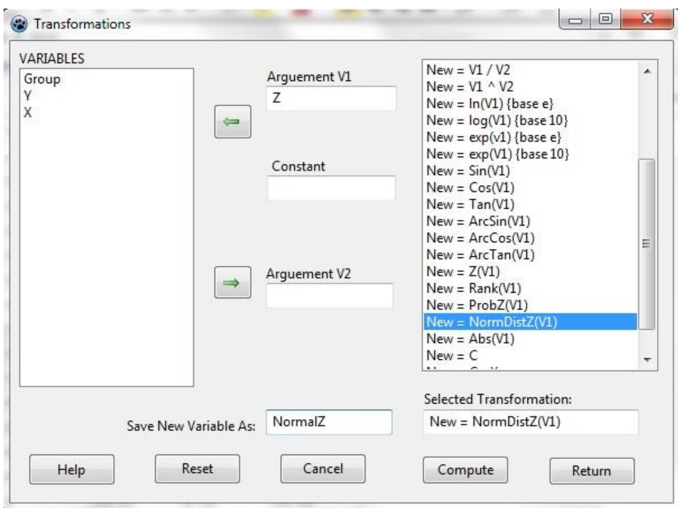


Fig. 1.5   The Transformation Form

You will note that you can transform a variable by adding, subtracting, multiplying, dividing or raising a value to a power. To do this you select a variable to transform by clicking on the variable in the list of available variables and then clicking the right arrow. You then enter a constant by clicking on the box for the constant and entering a value. You select the transformation with a constant from among the transformations by clicking on the desired transformation (you will see it entered automatically in the lower right box.) Next you enter a name for the new variable in the box labeled "Save new variable as:" and click the OK button.
Sometimes you will want to transform a variable using one of the common exponentiation or trigonometric functions. In this case you do not need to enter a constant - just select the variable, the desired transformation and enter the variable name before clicking the OK button.
You can also select a transformation that involves two variables. For example, you may want a new variable that represents the sum, product, difference, etc. of two variables. In this case you select the two variables for the first and second arguments using the appropriate right-arrow key after clicking one and then the other in the available variables list.

The "Print Definitions" option simply creates a list of variable definitions on an "output" form which may be printed on your printer for future reference.

The Enter An Equation option lets you create a new variable that combines existing variables with a variety of mathematical functions. The form below shows the form that appears when you select this option:

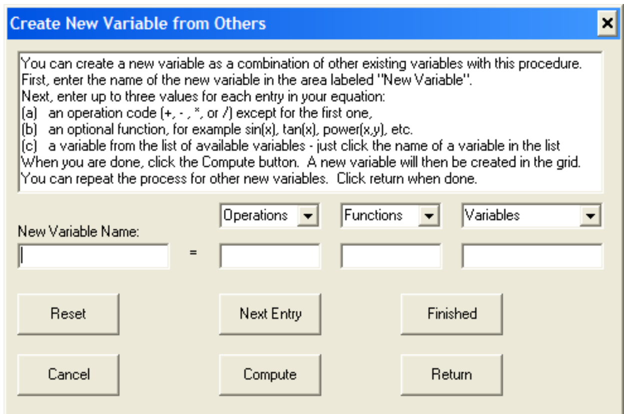


Fig. 1.6 The Equation Editor

This form lets you create an equation for a new variable such as:
NewVar = SQRT(Var1) * Log(Var3) – Var4
Typically, you will first enter the name for a new variable and then enter function or simply a variable from the drop-down box and then click the Next Entry button. The next entry will contain an operation, optionally a function and another variable from the Variables drop-down list. Continue this “Next Entry” process for the number of variables in your equation. When finished, click the “Compute” button to create the new variable values. Click the Return button to go back to the main form.

The Edit Menu

The Edit menu is provided primarily for deleting, cutting and pasting of cells, rows or columns of data. It also provides the ability to insert a new column or row at a desired position in the data grid. There is one special "paste" operation provided for users that also have the Microsoft Excel program and wish to copy cells from an Excel spreadsheet into the LazStats grid. These operations involve clicking on a cell in a given row and column and the selecting the edit operation desired. The user is encouraged to experiment with these operations in order to become familiar with them.

The Tools Menu

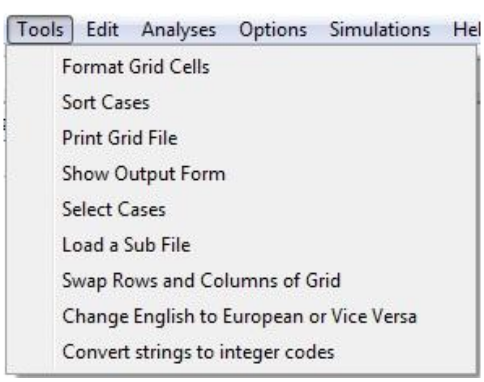


Fig. 1.7 The Tools Menu

An option under the Tools menu is to sort your data cases into ascending or descending sequence based on one of the selected variables. Shown below is the dialog for sorting cases:

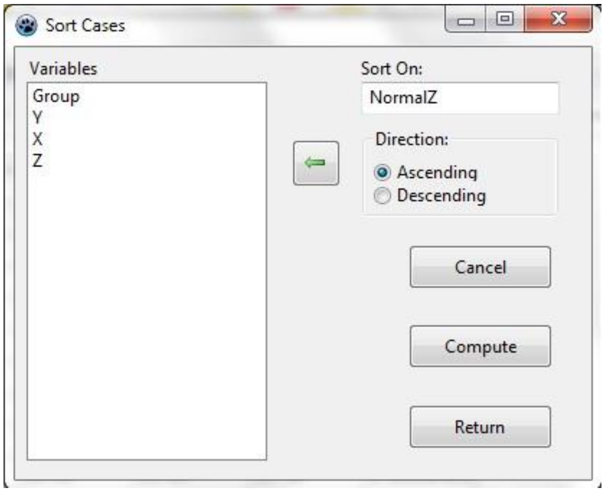


Fig. 1.8 The Sort Cases Menu

Another option under the Tools menu lets you switch between the American and European format for decimal fractions. This may be useful when you have imported a file from another country that uses the other format. LazStats will attempt to convert commas to periods or vice-versa as required.

You can open the “Output” form that is used to display results from the different procedures. The Output form is actually a minimal word processing procedure. You can write and edit text, change fonts, change colors, etc. as in many word processors. This will become advantageous as you will sometimes want to edit the output from a procedure to further enhance the results of an analysis prior to submission for publication.

The "Select Cases" option lets you analyze only those cases (rows) which you select. When you press this option you will see the following dialogue form:

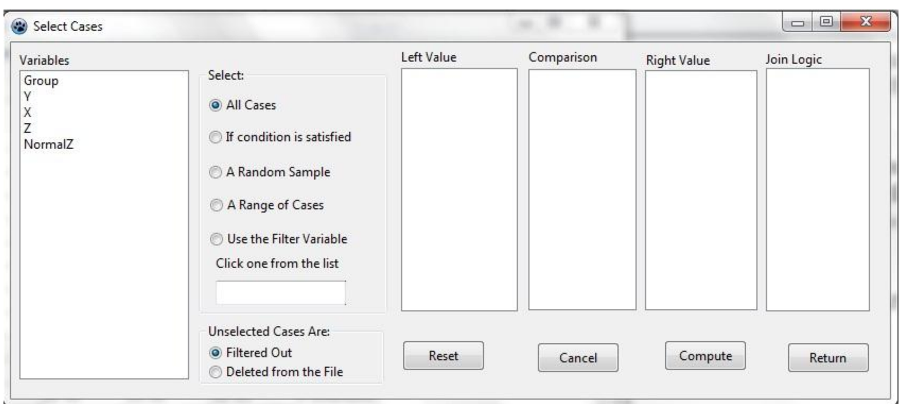


Fig. 1.9 The Select Cases Dialog Form

Notice that you may select a random number of cases, cases that exhibit a specific range of values or cases if a specific condition exists. Once selection has been made, a new variable is added to the grid called the "Filter" variable. You can subsequently use this filter variable to delete unneeded cases from your file if desired. Each of the selection procedures invokes a dialogue form that is specific to the type of selection chosen. For example, if you select the "if condition is satisfied" button, you will see the following dialogue form:

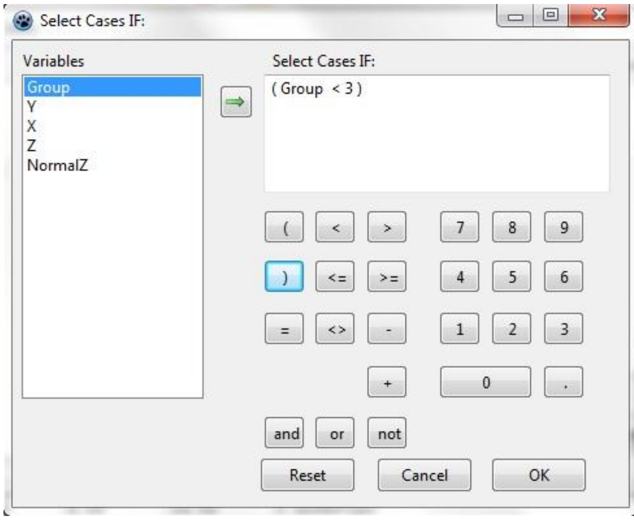


Fig. 1.10   The Select IF Dialog Form

An example has been entered on this form to demonstrate a typical selection criteria. Notice that compound statements involve the use of opening and closing parentheses around each expression and the combined expressions. You can directly enter values in the "if" box or use the buttons provided on the pad.

Should you select the "random" option in Fig. (1.9) you would see the following form:

Fig. 1.11   Selection of Cases at Random

The user may select a percentage of cases or select a specific number from a specified number of cases.

Finally, the user may select a specified range of cases. This option produces the following dialogue form:

Fig. 1.12   Selecting a Range of Cases

The Variables menu “Recode” option is used to change the value of cases in a given variable. For example, you may have imported a file that originally coded gender as "M" or "F" but the analysis you want requires a coding of 0 and 1. You can select the recode option and get the following form to complete:

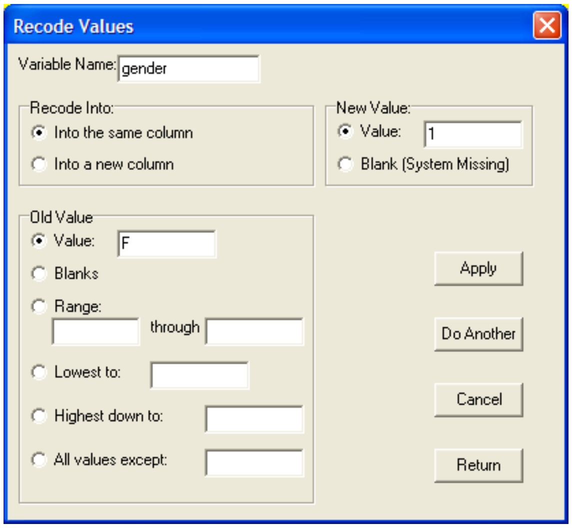


Fig. 1.13 Re-coding Values of a Variable

Notice that you must first click on the column of the variable to recode in the grid. Once you select the recode option, enter the old value (or value range) and also enter the new value before clicking the Apply button. You can repeat the process for multiple old values before returning to the Main Form.

The Analyses Menu

The heart of any statistics package is the ability to perform a variety of statistical analyses. Many of the typical analyses are included in the options and sub-options of the Analyses menu. The Fig. below shows the options and the sub-options under the descriptive option. No attempt will be made at this point in the text to describe each analysis - these are described further in the text.

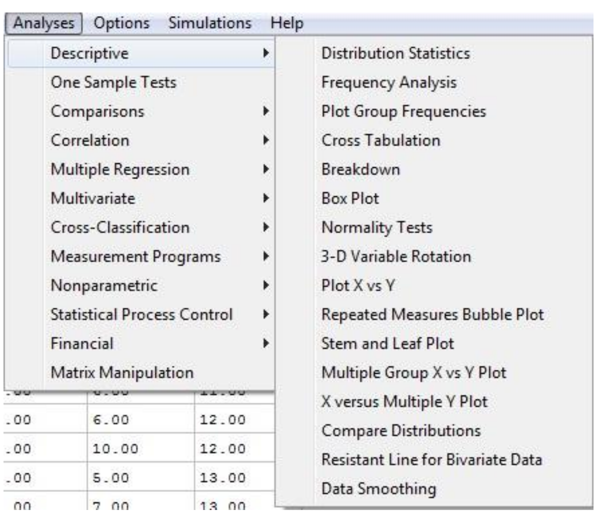


Fig. 1.14 The Analyses Menu, Descriptive Options

The Simulation Menu

As you read about and learn statistics, it is helpful to be able to simulate data for an analysis and see what the distribution of the values looks like. In addition, the concepts of "type I error", "type II error", "Power", correlation, etc. may be more readily grasped if the student can "play" with distributions and the effects of choices they might make in a real study. Under the simulation menu the user may generate a sequence of numbers, may generate multivariate data, may generate data that are a sample from a theoretical population or generate bivariate-normal data for a correlation. One can even generate data for a two-way analysis of variance! The Fig. below illustrates the Simulation Menu.

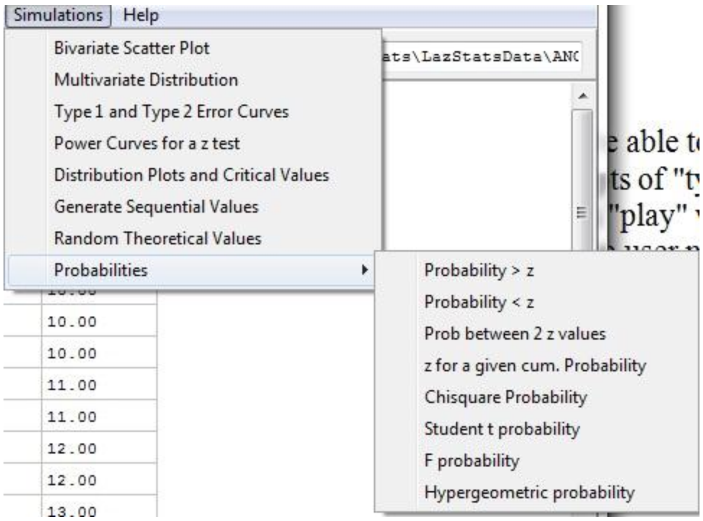


Fig. 1.15 The Simulation Menu

These simulation procedures are described later.

Creating Research Reports

Introduction

Once you begin using LazStats you may find it useful for creating research reports and articles. We will assume that you are a Windows operating system user and that you have installed on your computer a word processing package such as the free Open Office program or the Microsoft Word program. Once you begin creating your research document there are likely to be images from LazStats that you would like to include in that document. There are several ways to complete this task that we will cover in the following paragraphs. We assume that you will have started both LazStats and your word processing program so that you can switch between them as needed.

The Output Form

When you complete an analysis or simulation with LazStats, the printed output is placed on an “output form”. This output can be saved to your disk with a name that you choose. The output is saved as a “rich-text file”. This is a format that can be read by both of the previously mentioned software programs and those programs can “insert” another file into a currently opened file.

As an alternative, when the output form is shown by an LazStats procedure, you can drag your mouse over selected output while holding down the mouse button. The selected output will be highlighted as you do this. Copy the highlighted text to the Windows “Clipboard” by pressing the control key and the C key concurrently (Ctrl-C). Next, select the position in your document that you want to place the copied material by clicking on that position. Enter the Ctrl-V (concurrently press the control key and the V key) to copy the information on the Windows clipboard into your document.

Graphic Images

LazStats includes a variety of procedures that produce graphic pictures of data such as X versus Y plots, frequency distributions, power curves, etc.. Most of these graphic images can be saved to your disk and later included in your research document. They are saved as bitmap images (.bmp) files by LazStats and can be inserted into your word processing document where desired.

One can also click on any image displayed on your screen and press the Alternate key and the print screen keys concurrently to capture that image. You then use the paste function in your word processing program to transfer the saved image to your document. These images can be re-sized in your document. As a quick demonstration, I will click the alt-prtscr key combination on this current word document page and then paste in below using the ctrl-V key combination:

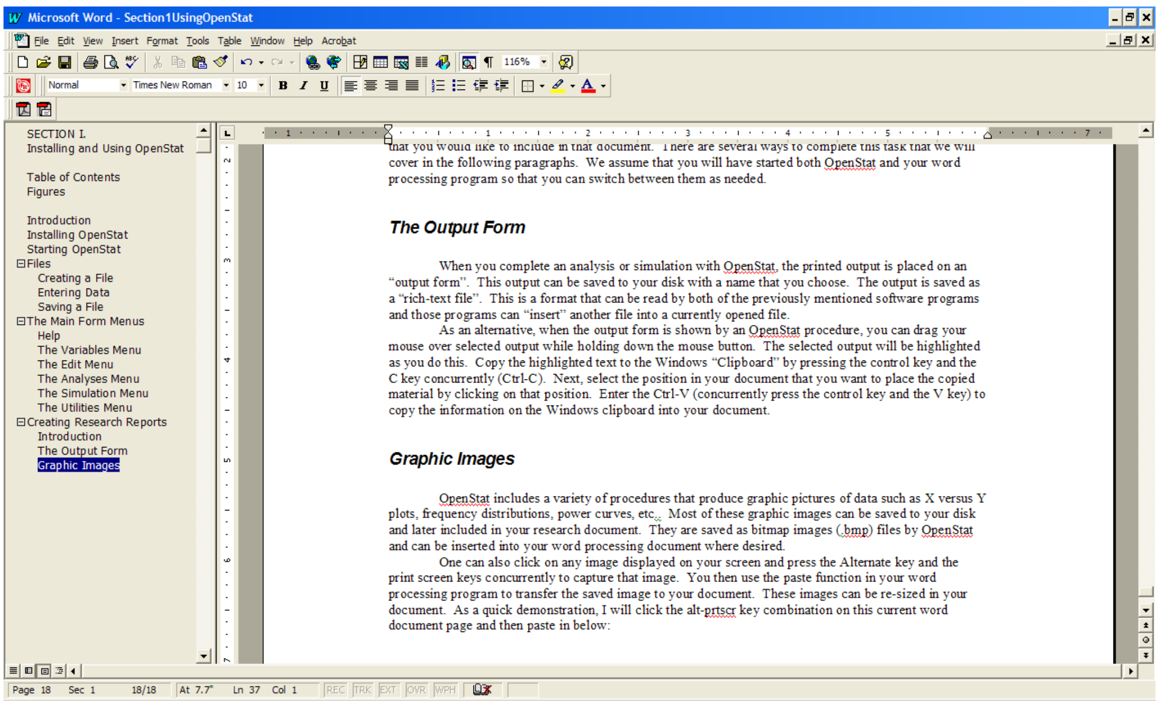


Fig. 1.16 Copying An Image Into A Document

Some Common Errors!

Empty Cells

The beginning user will often see a message something like "" is not a valid floating point value. The most common cause of this error occurs when a procedure attempts to read a blank cell, that is, a cell that has been left empty by the user. The new user will typically use the down-arrow to move to the next row in the data grid in preparation to enter the next row of values. If you do this after entering the values for the last case, you will create a row of empty cells. You should put the cursor on one of these empty cells and use the Edit->Delete Row menu to remove this blank row.

The user should define the "Missing Value" for each variable when they define the variable. One should also click on the Options menu and place a missing value in that form. LazStats attempts to place that missing value in empty cells when a file is saved as .LAZ file. Not all LazStats procedures allow missing values so you may have to delete cases with missing values for those procedures.

Incorrect Format for Floating Point Values

A second reason you might receive a "not valid" error is because you are using the European standard for the format of values with decimal fractions. Most of the statistical procedures contain a small "edit" window that contains a confidence level or a rejection area such as 95.0 or 0.05. These will NOT be valid floating point values in the European standard and the user will need to click on the value and replace it with the correct form such as 95,0 or 0,05. This has been done for the user in some procedures but not all!

String labels for Groups

Users of other statistics packages such as SPSS or Excel may have used strings of characters to identify different groups of cases (subjects or observations.) LazStats uses sequential integer values only in statistical analyses such as analyses of variance or discriminant function analysis. An edit procedure has been included that permits the conversion of string labels to integer values and saves those integers in a new column of the data grid. An attempt to use a string (alphanumeric) value will cause an "not valid" type of error. Several procedures in LazStats have been modified to let you specify a string label for a group variable and automatically create an integer value for the analysis in a few procedures but not all. It is best to do the conversion of string labels to integers and use the integer values as your group variable.

Floating Point Errors

Sometimes a procedure will report an error of the type "Floating Point Division Error". This is often the outcome of a procedure attempting to divide a quantity by zero (0.) As an example, assume you have entered data for several variables obtained on a group of subjects. Also assume that the value observed for one of those variables is the same (a constant value) for all cases. In this situation there is no variability among the cases and the variance and standard deviation will be zero! Now an attempt to use that zero variance or standard deviation in the calculation of z scores, a correlation with another variable or other usage will cause an error (division by zero is not defined.)

Values too Large (or small)

In some fields of study such as astronomy the values observed may be very, very large. Computers use binary numbers to represent quantities. Nearly all LazStats procedures use "double precision" storage for floating point values. The double precision value is stored in 64 binary "bits" in the computer memory. In most computers this is a combination of 8 binary "bytes" or words. The values are stored with a characteristic and mantissa similar to a scientific notation. Of course bits are also used to represent the sign of these parts. The maximum value for the characteristic is typically something like 2 raised to the power of 55 and the mantissa is 2 to the 7th power. Now consider a situation where you are summing the product of several of very large values such as is done in obtaining a variance or correlation. You may very well exceed the 64 bit storage of this large sum of products! This causes an "overflow" condition and a subsequent error message. The same thing can be said of values too small. This can cause an "underflow" error and associated error message.

The solution for these situations of values too large or too small is to "scale" your initial values. This is typically done by dividing or multiplying the original values by a constant to move the decimal point to decrease (or increase) the value. This does, of course, affect the "precision" of your original values but it may be a sacrifice necessary to do the analysis. In addition, the results will have to be "re-scaled" to reflect the original measurement scale.

Chapter 2.   Basic Statistics

Introduction

This chapter introduces the basic statistics concepts you will need throughout your use of the LazStats package. You will be introduced to the symbols and formulas used to represent a number of concepts utilized in statistical inference, research design, measurement theory, multivariate analyses, etc. Like many people first starting to learn statistics, you may be easily overwhelmed by the symbols and formulas - don't worry, that is pretty natural and does NOT mean you are retarded! You may need to re-read sections several times however before a concept is grasped. You will not be able to read statistics like a novel (don't we wish we could) but rather must "study" a few lines at a time and be sure of your understanding before you proceed.

Symbols Used in Statistics

Greek symbols are used rather often in statistical literature. (Is that why statistics is Greek to so many people?) They are used to represent both arithmetic types of operations as well as numbers, called parameters, that characterize a population or larger set of numbers. The letters you usually use, called Arabic letters, are used for numbers that represent a sample of numbers obtained from the population of numbers.

Two operations that are particularly useful in the field of statistics that are represented by Greek symbols are the summation operator and the products operator. These two operations are represented by the capital Greek letters Sigma Σ and Pi Π. Whenever you see these symbols you must think:

Σ= "The sum of the values: " , or
Π = "The product of the values:"

For example, if you see Y = Σ (1,3,5,9) you would read this as "the sum of 1, 3, 5 and 9". Similarly, if you see Y = Π(1,3,5,9) you would think "the product of 1 times 3 times 5 times 9".

Other conventions are sometimes adopted by statisticians. For example, as in beginning algebra classes, we often use X to represent any one of many possible numbers. Sometimes we use Y to represent a number that depends on one or more other numbers X_1, X_2, etc. Notice that we used subscripts of 1, 2, etc. to represent different (unknown) numbers. Lower case letters like y, x, etc. are also sometimes used to represent a deviation of a score from the mean of a set of scores. Where it adds to the understanding, X, and x may be italicized or written in a script style.

Now lets see how these symbols might be used to express some values. For example, we might represent the set of numbers (1,3,7,9,14,20) as X_1, X_2, X_3, X_4, X_5, and X_6. To represent the sum of the six numbers in the set we could write:

$$Y = \sum_{i=1}^{6} X_i = 1 + 3 + 7 + 9 + 14 + 20 = 54 \qquad (2.1)$$

If we want to represent the sum of any arbitrary set of N numbers, we could write the above equation more generally, thus

$$Y = \sum_{i=1}^{N} X_i \qquad (2.2)$$

represents the sum of a set of N values. Note that we read the above formula as "Y equals the sum of X subscript i values for the value of i ranging from 1 through N, the number of values".

What would be the result of the formula below if we used the same set of numbers (1,3,7,9,14,20) but each were multiplied by five ?

$$Y = \sum_{i=1}^{N} 5X_i = 5\sum_{i=1}^{N} X_i = 270 \tag{2.3}$$

To answer the question we can expand the formula to

$$\begin{aligned} Y &= 5X_1 + 5X_2 + 5X_3 + 5X_4 + 5X_5 + 5X_6 \\ &= 5(X_1 + X_2 + X_3 + X_4 + X_5 + X_6) \\ &= 5(1 + 3 + 7 + 9 + 14 + 20) \\ &= 5(54) = 270 \end{aligned} \tag{2.4}$$

In other words,

$$Y = \sum_{i=1}^{N} CX_i = C\sum_{i=1}^{N} X_i \tag{2.5}$$

We may generalize multiplying any sum by a constant (C) to

$$Y = \sum_{i=1}^{N} CX_i = C\sum_{i=1}^{N} X_i \tag{2.6}$$

What happens when we sum a term which is a compound expression instead of a simple value? For example, how would we interpret

$$Y = \sum_{i=1}^{N} (X_i - C) \tag{2.7}$$

where C is a constant value?

We can expand the above formula as

$$Y = (X_1 - C) + (X_2 - C) + ... + (X_N - C) \tag{2.8}$$

(Note the use of ... to denote continuation to the N^{th} term).

The above expansion could also be written as

$$Y = (X_1 + X_2 + ... + X_N) - NC \tag{2.9}$$

Or $Y = \sum_{i=1}^{N} X_i - NC$ (2.10)

We note that the sum of an expression which is itself a sum or difference of multiple terms is the sum of the individual terms of that expression. We may say that the summation operator distributes over the terms of the expression!

Now lets look at the sum of an expression which is squared. For example,

$$Y = \sum_{i=1}^{N} (X_i - C)^2 \tag{2.11}$$

When the expression summed is not in its most simple form, we must first evaluate the expression. Thus

$$Y = \sum_{i=1}^{N}(X_i - C)^2 = \sum_{i=1}^{N}(X_i - C)(X_i - C) = \sum_{i=1}^{N}\left[X_i^2 - 2CX_i + C^2\right] = \sum_{i=1}^{N}X_i^2 - \sum_{i=1}^{N}2CX_i + \sum_{i=1}^{N}C^2 \tag{2.12}$$

or $Y = \sum_{i=1}^{N}X_i^2 - 2C\sum_{i=1}^{N}X_i + NC^2 = \sum_{i=1}^{N}X^2 - 2CN\overline{X} - NC^2 = \sum_{i=1}^{N}X^2 - CN(2\overline{X} - C)$

(2.12)

Probability Concepts

Maybe, possibly, could be, chances are, probably are all words or phrases we use to convey uncertainty about something. Yet all of these express some belief that a thing or event could occur or exist. The field of statistics is concerned about making such statements based on observations that will lead us to correct "guesses" about an event occuring or existing. The field of study called "statistics" gets its name from the use of samples that we can observe to estimate characteristics about the population that we cannot observe. If we can study the whole population of objects or events, there is no need for statistics! Accounting methods will suffice to describe the population. The characteristics (or indexes) we observe about a sample from a population are called *statistics*. These indexes are estimates of population characteristics called *parameters*. It is the job of the statistician to provide indexes (statistics) about populations that give us some level of *confidence* that we have captured the true characteristics of the population of interest.

When we use the term *probability* we are talking about the *proportion* of objects in some population. It might be the proportion of some discrete number of heads that we get when tossing a coin. It might be the proportion of values within a specific range of values we find when we observe test scores of student achievement examinations.

In order for the statistician to make useful observations about a sample that will help us make confident statements about the population, it is often necessary to make *assumptions* about the *distribution* of scores in the population. For example, in tossing a coin 30 times and examining the outcome as the number of heads or tails, the statistician would assume that the distribution of heads and tails after a very large number of tosses would follow the *binomial* distribution, a theoretical distribution of scores for a binary object. If the population of interest is the relationship between beginning salaries and school achievement, the statistician may have to assume that the measures of salary and achievement have a *normal* distribution and that the relationship can be described by the *bivariate-normal* distribution.

A variety of indexes (statistics) have been developed to estimate characteristics (measurements) of a population. There are statistics that describe the *central tendency* of the population such as the mean (average), median and mode. Other statistics are used to describe how variable the scores are. These statistics include the variance, standard deviation, range, semi-interquartile range, mean deviation, etc. Still other indices are used to describe the relationship among population characteristics (measures) such as the product-moment correlation and the multiple regression coefficient of determination. Some statistics are used to examine differences among samples from possibly different populations to see if they are more likely to be samples from the same population. These statistics include the "t" and "z" statistic, the chi-squared statistic and the F-Ratio statistic.

The sections below will describe many of the statistics obtained on samples to make inferences about population parameters. The assumed (theoretical) distribution of these statistics will also be described.

Additive Rules of Probability

Formal aspects of probability theory are discussed in this section. But first, we need to define some terms we will use. First, we will define a *sample space* as simply a set of points. A point can represent anything like persons, numbers, balls, accidents, etc. Next we define an *event*. An event is an observation of something happening such as the appearance of "heads" when a coin is tossed or the observation that a person you selected at random from a telephone book is voting Democrat in the next election. There may be several points in the sample space, each of which is an example of an event. For example, the sample space may consist of 5 black balls and 4 white balls in an urn. This sample space would have 9 points. An event might be "a ball is black." This event has 5 sample space points. Another event might be "a ball is white." This event has a sample space of 4 points. We may now say that the probability of an event E is the ratio of the number of sample points that are examples of E to the total number of sample points provided all sample points are equally likely. We will use the notation P(E) for the probability of an event. Now let an event be "A ball is black" where the sample space is the set of 9 balls (5 black and 4 white.) There are 5 sample points that are examples of this event out of a total of 9 sample points. Thus the probability of the event $P(E) = 5 / 9$. Notice that the probability that a ball is white is 4/9. We may also say that the probability that a ball is red is 0 / 9 or that the probability that the ball is both white and black is 0 / 9. What is the probability that the ball is either white OR black? Clearly this is $(5 + 4) / 9 = 1.0$.

In our previous example of urn balls, we noticed that a ball is either white or black. These are mutually exclusive events. We also noted that the sum of exclusive events is 1.0. Now let us add 3 red balls to our urn. We will label our events as B, W or R for the colors they represent. Our sample space now has 12 points. What is the probability that two balls selected are either B or W? When the events are exclusive we may write this as $P(B \cup A)$. Since these are exclusive events, we can write: $P(B \cup W) = P(B) + P(W) = 5 / 12 + 4 / 12 = 9 / 12 = 3 / 4 = 0.75$.

It is possible for a sample point to be an example of two or more events. For example if we toss a "fair" coin three times, we can observe eight possible outcomes:

1. HHH   2. HHT   3. HTH   4. HTT   5. TTT   6. TTH   7. THT and 8. THH

If our coin is fair we can assume that each of these outcomes is equally likely, that is, has a probability of 1/8. Now let us define two events: event A will be getting a "heads" on flip 1 and flip 2 of the coin and event B will be getting a "heads" on flips 1 and 3 of the coin. Notice that outcomes 1 and 2 above are sample points of event A and that outcomes 1 and 3 are events of type B. Now we can define a new event that combines events A *and* B. We will use the symbol $A \cap B$ for this event. If we assume each of the eight sample points are equally likely we may write $P(A \cap B)$ = number of sample points that are examples of $A \cap B$ / total number of sample points, or
$P(A \cap B) = 1 / 8$. Notice that only 1 of the points in our sample space has heads on both flips 1 and 2 and on 2 and 3 (sample point 1.) That is, the probability of event A *and* B is the probability that both events A and B occur.

When events may not be exclusive, we are dealing with the probability of an event A or Event B or both. We can then write

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (2.13)$$

Which, in words says, the probability of events A and B equals the probability of event A plus the probability of event B minus the probability of event A and B. Of course, if A and B are mutually exclusive then the probabilty of A and B is zero and the probability of A or B is simply the sum of P(A) and P(B).

The Law of Large Numbers

Assume again that you have an urn of 5 black balls and 4 white balls. You stir the balls up and draw one from the urn and record its color. You return the ball to the urn, again stir the balls vigourously and again draw a single ball and record its color. Now assume you do this 10,000 times, each time recording the color of the ball. Finally, you count the number of white balls you drew from the 10,000 draws. You might reasonably expect the proportion of white balls to be close to 4/9 although it is likely that it is not exactly 4/9. Should you continue to repeat this experiment over and over, it is also reasonable to expect that eventually, the proportion would be extremely close to the actual proportion of 4/9. You can see that the larger the number of observations, the more closely we would approximate the actual value. You can also see that with very small replications, say 12 draws (with replacement) could lead to a very poor estimate of the actual proportion of white balls.

Multiplication Rule of Probability

Assume you toss a fair coin five times. What is the probability that you get a "heads" on all five tosses? First, the probability of the event P(E) = 1/2 since the sample space has only two possible outcomes. The multicative rule of probability states that the probability of five heads would be 1/2 * 1/2 * 1/2 * 1/2 * 1/2 or simply (1/2) to the fifth power (1/32) or, in general, $P(E)^n$ where n is the number of events E.

As another example of this rule, assume a student is taking a test consisting of six multiple-choice items. Each item has 5 equally attractive choices. Assume the student has absolutely no knowledge and therefore guesses the answer to each item by randomly selecting one of the five choices for each item. What is the probability that the student would get all of the items correct? Since each item has a probability of 1/5, the probability that all items are answered correctly is $(1/5)^6$ or 0.000064 . What would it be if the items were true-false items?

Permutations and Combinations

A *permutation* is an arrangement of n objects. For example, consider the letters A, B, C and D. How many permutations (arrangements) can we make with these four letters? We notice there are four possibilities for the first letter. Once we have selected the first letter there are 3 possible choices for the second letter. Once the second letter is chosen there are two possibilities for the third letter. There is only one choice for the last letter. The number of permutations possible then is 4 x 3 x 2 x 1 = 24 ways to arrange the four letters. In general, if there are N objects, the number of permutations is N x (N-1) x (N-2) x (N-3) x … (1). We abbreviate this series of products with an exclamation point and write it simply as N! We say "N factorial" for the product series. Thus 4! = 24. We do, however, have to let 0! = 1, that is, by definition the factorial of zero is equal to one. Factorials can get very large. For example, 10! = 3,628,800 arrangements. If you spent a minute examining one arrangement of 12 guests for a party, how long would it take you to examine each arrangement? I'm afraid that if you worked 8 hours a day, five days a week for 52 weeks a year you (and your descendants) would still be working on it for more than a thousand years!

A *combination* is a set of objects without regard to order. For example, the combination of A, B, C and D in any permutation is one combination. A question arises however concerning how many combinations of K objects can be obtained from a set of N objects. For example, how many combinations of 2 objects can be obtained from a set of 4 objects. In our example, we have the possibilities of A + B, A + C, A + D, B + C, B + D and C + D or a total of 6 combinations. Notice that the combination AB is the same as BA because order is not considered. A formula may be written using permutations that gives us a general formula for combinations. It is

$$N! / [\, K!\, (N-K)!] \tag{2.14}$$

In our example then, the number of combinations of 2 things out of 4 is 4! / [2! (4-2)!] which might be written as

$$\frac{4 \times 3 \times 2 \times 1}{(2 \times 1) \times (2 \times 1)} = \frac{24}{4} = 6 \tag{2.15}$$

A special mathematics notation is often used for the combination of k things out of N things. It is

$$\binom{N}{K} = \frac{N!}{K!(N-K)!} \tag{2.16}$$

You will see the use of combinations in the section on the binomial distribution.

Conditional Probability

In sections above we defined the additive law for mutually exclusive events as the sum of the invidual probabilities. For example, for a fair die the probability of each of the faces is 1/6 so the probability of getting a 1 in two tosses (toss A and a toss B) is P(A) + P(B) = 1/6 + 1/6 = 1/3. Our multiplicative law for independent events states that the probability of obtaining event A *and* event B is P(A) x P(B). So the probability of getting a 1 on toss A of a die 1 *and* toss B of the die is P(1) x P(2) = 1/6 x 1/6 = 1/36. But what if we don't know our die is a "fair" die with equal probabilties for each face on a toss? Can we use the prior information from toss A of the die to say what the probability if for toss B?

Conditional probability is the probability of an event given that another event has already occurred. We would write

$$P(B \mid A) = \frac{P(A \bigcap B)}{P(A)} \tag{2.17}$$

If A and B are independent then

$$P(B \mid A) = \frac{P(A)P(B)}{P(A)} = P(B) \tag{2.18}$$

or the probability of the second toss is 1/6, the same as before.

Now consider two events A and B: for B an individual has tossed a die four times with outcomes E1, E2, E3 and E4; For A the event is the tosses with outcomes E1 and E2. The events might be the toss results of 1, 3, 5 and 6. Knowing that event A has occurred, what is the probabilty of event B, that is, P(A|B)? Intuitively you might notice that the probabilty of the B event is the sum of the individual probabilities or 1/6 + 1/6 + 1/6 + 1/6 = 2/3, and that the probability of the A event is 1/6 + 1/6 = 1/3 or half the probability of B. That is, P(A) / P(B) = 1/2.

A more formal statement of conditional probability is

$$P(A \mid B) = \frac{P(A \bigcap B)}{P(B)} \tag{2.19}$$

Thus the probability of event A is conditional on the prior probability of B. The result P(A|B) is sometimes called the posterior probability. Notice we can rewrite the above equation as:

$$P(A \mid B)P(B) = P(A \bigcap B) \qquad \{2.20)$$

and

$$P(B \mid A)P(A) = P(A \bigcap B) \tag{2.21}$$

Since both equations equal the same thing we may write

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.22}$$

The above is known as Bayes Theorem for events.

Now consider an example. In a recent poll in your city, 40 percent are registered Democrats and 60 percent are registered Republicans. Among the Democrats, the poll shows that 70% feel that invading Iraq was a mistake and 20% feel it was justified. You have just met a new neighbor and have begun a conversation over a cup of coffee. You learn that this neighbor feels that invading Iraq was a mistake. What is the probability that the neighbor is also a Democrat? Let A be the event that the neighbor is Democrat and B be the event that she feels the invasion was a mistake. We already know that the probability of A is P(A) = 0.6. We also know that the probability of B is P(B|A) = 0.7 . We need to compute P(B), the probability the neighbor feels the invasion was a mistake. We notice that the probability of B can be decomposed into two exclusive parts: P(B) = P(B and A) and P(B and *not* A) where the probability of *not* A is 1 - P(A) or 0.4, the probability of not being a democrat. We can write

$$P(B \bigcap notA) = P(notA)P(B \mid A) \tag{2.23}$$

or P(B) = P(B and A) + P(*not* A)P(B|*not* A) (2.24)

or P(B) = P(B|A)P(A) + P(*not* A) P(B|*not* A) (2.25)

Now we know P(A) = 0.4, P(*not* A) = 1 - .4 = 0.6, P(B|A) = 0.7 and P(B| *not* A) = 0.2. Therefore,
P(B) = (0.7) (0.4) + (0.6)(0.2) = 0.40

Now knowing P(B) we can compute P(A|B) using Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{(0.7)(0.4)}{0.4} = 0.7 \tag{2.26}$$

is the probability of the neighbor being Democrat.

Probabilty as an Area

Probabilities are often represented as proportions of a circle or a polygon that shows the distribution of events in a sample space. Venn diagrams are circles with a portion of the ellipse shaded to represent a probability of an event in the space of the circle. In this case the circles area is considered to be 1.0. Distributions for binomial events, normally distributed events, poisson distributed events, etc. will often show a shaded area to represent a probability. You will see these shapes in sections to come.

Sampling

In order to make reasonable inferences about a population from a sample, we must insure that we are observing sample data that is not, in some artificial way, going to lead us to wrong conclusions about the population. For example, if we sample a group of Freshman college students about their acceptance or rejection of abortion, and use this to estimate the beliefs about the population of adults in the United States, we would not be collecting an *unbiased* or fair sample. We often use the term *experiment* to describe the process of drawing a sample. A *random experiment* or random sample is considered a fair or un-biased basis for estimating population parameters. You can appreciate the fact that the number of experiments (samples) drawn is highly critical to make relevant inferences about the population. For example, a series of four tosses of a coin and counting the number of heads that occur is a rather small number of samples from which to infer whether or not the coin is likely to yield 50% heads and 50% tails if you were to continue to toss the coin an infinite number of times! We will have much more confidence about our sample statistics if we use a large number of experiments.

Two of the most common mistakes of beginning researchers is failing to use a random sample and to use too few samples (observations) in their research. A third common mistake is to assume a theoretical model for the distribution of sample values that is incorrect for the population.

The Mean

The mean is probably the most often used parameter or statistic used to describe the central tendency of a population or sample. When we are discussing a population of scores, the mean of the population is denoted with the Greek letter μ . When we are discussing the mean of a sample, we utilize the letter X with a bar above it. The sample mean is obtained as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.27}$$

The population mean for a finite population of values may be written in a similar form as

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{2.28}$$

When the population contains an infinite number of values which are continuous, that is, can be any real value, then the population mean is the sum of the X values times the proportion of those values. The sum of values which can be an arbitrarily small in differences from one another is written using the integral symbol instead of the Greek sigma symbol. We would write the mean of a set of scores that range in size from minus infinity to plus infinity as

$$\mu = \int_{-\infty}^{+\infty} Xp(X)dx \tag{2.29}$$

where p(X) is the proportion of any given X value in the population. The tall curve which resembles a script S is a symbol used in calculus to mean the "sum of" just like the symbol Σ that we saw previously. We use Σ to represent "countable" values, that is values which are discrete. The "integral" symbol on the other hand is used to represent the sum of values which can range continuously, that is, take on infinitely small differences from one-another.

A similar formula can be written for the sample mean, that is,

$$\overline{X} = \sum_{i=1}^{n} X_i p(X_i) \tag{2.30}$$

where p(X) is the proportion of any given X_i value in the sample.

If a sample of n values is randomly selected from a population of values, the sample mean is said to be an unbiased estimate of the population mean. This simply means that if you were to repeatedly draw random samples of size n from the population, the average of all sample means would be equal to the population mean. Of course we rarely draw more than one or two samples from a population. The sample mean we obtain therefore will typically *not* equal the population mean but will in fact differ from the population mean by some specific amount. Since we usually don't know what the population mean is, we therefore don't know how far our sample mean is from the population mean. If we have, in fact, used random sampling though, we do know something about the shape of the distribution of sample means; they tend to be *normally* distributed. (See the discussion of the Normal Distribution in the section on Distributions). In fact, we can estimate how far the sample mean will be from the population mean some (P) percent of the time. The estimate of sampling errors of the mean will be further discussed in the section on testing hypotheses about the difference between sample means.

Now let us examine the calculation of a sample mean. Assume you have randomly selected a set of 5 scores from a very large population of scores and obtained the following:

$X_1 = 3$
$X_2 = 7$
$X_3 = 2$
$X_4 = 8$
$X_5 = 5$

The sample mean is simply the sum ($\sum$) of the X scores divided by the number of the scores, that is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \sum_{i=1}^{5} (X_1 + X_2 + X_3 + X_4 + X_5)/5 = (3+7+2+8+5)/5 = 5.0 \tag{2.31}$$

We might also note that the proportion of each value of X is the same, that is, one out of five. The mean could also be obtained by

$$\overline{X} = \sum_{i=1}^{n} X_i \, p(X_i) \tag{2.32}$$

= 3 (1/5) + 7 (1/5) + 2 (1/5) + 8 (1/5) + 5 (1/5)

= 5.0

The sample mean is used to indicate that value which is "most typical" of a set of scores, or which describes the center of the scores. In fact, in physics, the mean is the center-of-gravity (sometimes called the first moment of inertia) of a solid object and corresponds to the fulcrum, the point at where the object is balanced.

Unfortunately, when the population of scores from which we are sampling is not symmetrically distributed about the population mean, the arithmetic average is often not very descriptive of the "central" score or most representative score. For example, the population of working adults earn an annual salary of $21,000.00. These salaries however are not symmetrically distributed. Most people earn a rather modest income while there are a few who earn millions. The mean of such salaries would therefore not be very descriptive of the typical wage earner. The mean value would be much higher than most people earn. A better index of the "typical" wage earner would probably be the *median*, the value which corresponds to the salary earned by 50 percent or fewer people.

Examine the two sets of scores below. Notice that the first 9 values are the same in both sets but that the tenth scores are quite different. Obtain the mean of each set and compare them. Also examine the score below which 50 percent of the remaining scores fall. Notice that it is the same in both sets and better represents the "typical" score.

SET A: (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Mean = ?
Median = ?

SET B: (1, 2, 3, 4, 5, 6, 7, 8, 9, 1000)

Mean = ?
Median = ?

Variance and Standard Deviation

A set of scores are seldom all exactly the same if they represent measures of some attribute that varies from person to person or object to object. Some sets of scores are much more variable that others. If the attribute measures are very similar for the group of subjects, then they are less variable than for another group in which the subjects vary a great deal. For example, suppose we measured the reading ability of a sample of 20 students in the third grade. Their scores would probably be much less variable than if we drew a sample of 20 subjects from across the grades 1 through 12!

There are several ways to describe the variability of a set of scores. A very simple method is to subtract the smallest score from the largest score. This is called the *exclusive range*. If we think the values obtained from our measurement process are really point estimates of a continuous variable, we may add 1 to the exclusive range and obtain the *inclusive range*. This range includes the range of possible values. Consider the set of scores below:

5, 6, 6, 7, 7, 7, 8, 8, 9

If the values represent discrete scores (not simply the closest value that the precision of our instrument gives) then we would use the exclusive range and report that the range is (9 - 5) = 4. If, on the other hand, we felt that the scores are really point estimates in the middle of intervals of width 1.0 (for example the score 7 is actually an observation someplace between 6.5 and 7.5) then we would report the range as (9-5) + 1 = 5 or (9.5 - 4.5) = 5.

While the range is useful in describing roughly how the scores vary, it does not tell us much about how MOST of the scores vary around, say, the mean. If we are interested in how much the scores in our set of data tend to differ from the mean score, we could simply average the distance that each score is from the mean. The mean deviation, unfortunately is always 0.0! To see why, consider the above set of scores again:

Mean = (5+6+6+7+7+7+8+8+9) / 9 = 63 / 9 = 7.0

Now the deviation of each score from the mean is obtained by subtracting the mean from each score:

5 - 7 = -2
6 - 7 = -1
6 - 7 = -1
7 - 7 = 0
7 - 7 = 0
7 - 7 = 0
8 - 7 = +1
8 - 7 = +1
9 - 7 = +2
____

Total = 0.0

Since the sum of deviations around the mean always totals zero, then the obvious thing to do is either take the average of the absolute value of the deviations OR take the average of the squared deviations. We usually average the squared deviations from the mean because this index has some very important application in other areas of statistics.

The average of squared deviations about the mean is called the *variance* of the scores. For example, the variance, which we will denote as S^2, of the above set of scores would be:

$$S^2 = \frac{(-2)^2 + (-1)^2 + (-1)^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 2^2}{9} = 1.3333 \quad \text{approximately.} \tag{2.33}$$

Thus we can describe the score variability of the above scores by saying that the average squared deviation from the mean is about 1.3 score points.

We may also convert the average squared value to the scale of our original measurements by simply taking the square root of the variance, e.g. $S = \sqrt{(1.3)} = 1.1547$ (approximately). This index of variability is called the *standard deviation* of the scores. It is probably the most commonly used index to describe score variability!

Estimating Population Parameters : Mean and Standard Deviation

We have already seen that the mean of a sample of scores randomly drawn from a population of scores is an estimate of the population's mean. What we have to do is to imagine that we repeatedly draw samples of size n from our population (always placing the previous sample back into the population) and calculate a sample mean each time. The average of all (infinite number) of these sample means is the population mean. In algebraic symbols we would write:

$$\mu = \frac{\sum_{i=1}^{k} \overline{X}_i}{k} \text{ as k} \rightarrow \infty \tag{2.34}$$

Notice that we have let $\overline{X}$ represent the sample mean and μ represent the population mean. We say that the sample mean is an *unbiased* estimate of the population mean because the average of the sample statistic calculated in the same way that we would calculate the population mean leads to the population mean. We calculate the sample mean by dividing the sum of the scores by the number of scores. If we have a finite population, we could calculate the population mean in exactly the same way.

The sample variance calculated as the average of squared deviations about the sample mean is, however, a *biased* estimator of the population variance (and therefore the standard deviation also a biased estimate of the population standard deviation). In other words, if we calculate the average of a very large (infinite) number of sample variances this average will NOT equal the population variance. If, however, we multiply each sample variance by the constant n / (n-1) then the average of these "corrected" sample variances will, in fact, equal the population variance! Notice that if n, our sample size, is large, then the bias n / (n-1) is quite small. For example a sample size of 100 gives a correction factor of about 1.010101. The bias is therefore approximately 1 hundredth of the population variance. The reason that the average of squared deviations about the sample means is a biased estimate of the population variance is because we have a slightly different mean (the sample mean) in each sample.

If we had knowledge of the population mean μ and always subtracted μ from our sample values X, we would not have a biased statistic. Sometimes statisticians find it more convenient to use the biased estimate of the population variance than the unbiased estimate. To make sure we know which one is being used, we will use different symbols for the biased and unbiased estimates. The biased estimate will be represented here by a S^2 and the unbiased by a s^2 . The reason for use of the square symbol is because the square root of the variance is the standard deviation. In other words we use S for the biased standard deviation and s for the unbiased standard deviation. The Greek symbol sigma σ is used to represent the population standard deviation and σ^2 represents the population variance. With these definitions in mind then, we can write:

$$\sigma^2 = \frac{\sum_{j=1}^{K} s_i^2}{k} \text{ as k} \rightarrow \infty \tag{2.35}$$

or

$$\sigma^2 = \frac{\sum_{j}^{k} \frac{n}{n-1} S_j^2}{k} \text{ as k} \rightarrow \infty \tag{2.36}$$

where n is the sample size, k the number of samples, S^2 is the biased sample variance and s^2 is the unbiased sample variance.

You may have already observed that multiplying the biased sample variance by n / (n-1) gives a more direct way to calculate the unbiased variance, that is:

$s^2 = (n / (n-1)) * S^2$ or

$$s^2 = \frac{n}{n-1} \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} \tag{2.37}$$

In other words, we may directly calculate the unbiased estimate of population variance by dividing the sum of square deviations about the mean by the sample size minus 1 instead of just the sample size.

The numerator term of the variance is usually just called the "sum of squares" as sort of an abbreviation for the sum of squared deviations about the mean. When you study the Analysis of Variance, you will see a much more extensive use of the sum of squares. In fact, it is even further abbreviated to SS . The unbiased variance may therefore be written simply as

$$s^2 = \frac{SS_x}{n-1}$$

The Standard Error of the Mean

In the previous discussion of unbiased estimators of population parameters, we discussed repeatedly drawing samples of size n from a population with replacement of the scores after drawing each sample. We noted that the sample mean would likely vary from sample to sample due simply to the variability of the scores randomly selected in each sample. The question may therefore be asked "How variable ARE the sample means?". Since we have already seen that the variance (and standard deviation) are useful indexes of score variability, why not use the same method for describing variability of sample means? In this case, of course, we are asking how much do the sample means tend to vary, on the average, around the population mean. To find our answer we could draw, say, several hundred samples of a given size and calculate the average of the sample means to estimate Since we have already seen that the variance (and standard deviation) are useful indexes of score variability, why not use the same method for describing variability of sample means? In this case, of course, we are asking how much do the sample means tend to vary, on the average, around the population mean. To find our answer we could draw, say, several hundred samples of a given size and calculate the average of the sample means to estimate μ and then get the squared difference of each sample mean from this estimate. The average of these squared deviations would give us an approximate answer. Of course, because we did not draw ALL possible samples, we would still potentially have some error in our estimate. Statisticians have provided mathematical proofs of a more simple, and unbiased, estimate of how much the sample mean is expected to vary. To estimate the variance of sample means we simply draw ONE sample, calculate the unbiased estimate of X score variability in the population then divide that by the sample size! In symbols

$$s^2_{\overline{X}} = \frac{s^2_X}{n} \tag{2.38}$$

The square root of this estimate of variance of sample means is the estimate of the standard deviation of sample means. We usually refer to this as the *standard error of the mean*. The standard error of the mean represents an estimate of how much the means obtained from samples of size n will tend to vary from sample to sample. As an example, let us assume we have drawn a sample of 7 scores from a population of scores and obtained :

1, 3, 4, 6, 6, 2, 5

First, we obtain the sample mean and variance as :

$$\overline{X} = \frac{\sum_{i=1}^{7} X_i}{7} = 3.857 \text{ (approximately)} \tag{2.39}$$

$$s^2 = \frac{\sum_{i=1}^{7}\left(X_i - \overline{X}\right)^2}{7-1} = \frac{127}{6} = 3.81 \tag{2.40}$$

Then the variance of sample means is simply

$$s_{\overline{X}}^2 = \frac{s_X^2}{n} = \frac{3.81}{7} = 0.544 \qquad (2.41)$$

and the standard error of the mean is estimated as

$$s_{\overline{X}} = \sqrt{s_{\overline{X}}^2} = 0.74 \qquad (2.42)$$

You may have noticed by now, that as long as we are estimating population parameters with sample statistics like the sample mean and sample standard deviation, that it is theoretically possible to obtain estimates of the variability of ANY sample statistic. In principle this is true, however, there are relatively few that have immediate practical use. We will only be using the expected variability of a few sample statistics. As we introduce them, we will tell you what the estimate is of the variance or standard deviation of the statistic. The standard error of the mean, which we just examined, will be used in the z and t-test statistic for testing hypotheses about single means. More on that later..

Descriptive Statistics With LazStats

This section demonstrates the use of LazStats to obtain descriptive statistics for data that you have entered in a file on the main form's grid. In many cases, a graphical picture of one's data is highly useful in understanding the distribution of the values for one or more variables. In some procedures, the data of one or more variables must be defined as an integer. In other procedures, the data should be defined as a floating point variable. Be sure to define your variables as needed for each procedure.

Central Tendency and Variability

Click on the Analyses menu and place your mouse on the Descriptive option. The sub-option for Distribution Statistics is then chosen by clicking that option. To demonstrate, we will use the file labeled cansas.LAZ and obtain the descriptive statistics for the variable "Weight".

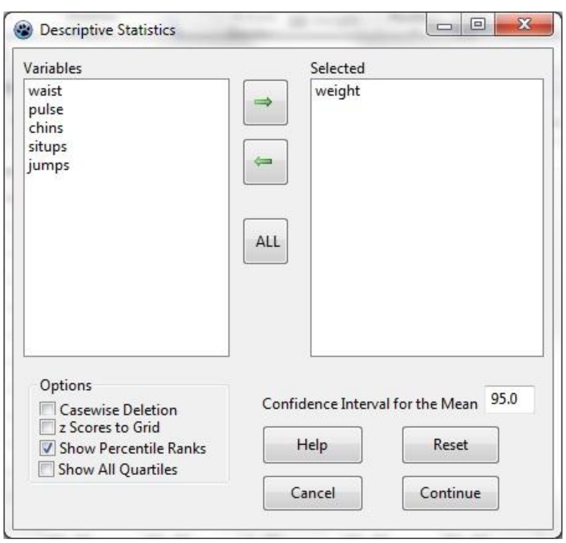


Fig. 2.1 The Dialog for Central Tendency and Variability

When you click the Continue button on the above form you will see the output displayed on the "Output Form". Notice that there are several options that may have been selected. The CaseWise Deletion option lets you obtain the results for only those cases in which there are no missing values. The z Scores to Grid option lets you create new variables that are the standardized z scores (mean of 0 and standard deviation of 1.0) for the variables you selected to analyze. Shown below is the result of our analysis:

```
DISTRIBUTION PARAMETER ESTIMATES

weight (N = 20)  Sum =       3572.000
Mean =    178.600  Variance =    609.621  Std.Dev. =     24.691
Std.Error of Mean =      5.521
95.00 percent Confidence Interval for the mean =  167.083 to  190.117
Range =    109.000  Minimum =    138.000  Maximum =    247.000
Skewness =      0.970  Std. Error of Skew =      0.512
Kurtosis =      1.802  Std. Error Kurtosis =      0.992

PERCENTILE RANKS
Score Value   Frequency   Cum.Freq. Percentile Rank

___________   __________  __________ ______________
   138.000          1           1        2.50
   154.000          2           3       10.00
   156.000          1           4       17.50
   157.000          1           5       22.50
   162.000          1           6       27.50
   166.000          1           7       32.50
   167.000          1           8       37.50
   169.000          1           9       42.50
   176.000          2          11       50.00
   182.000          1          12       57.50
   189.000          2          14       65.00
   191.000          1          15       72.50
   193.000          2          17       80.00
   202.000          1          18       87.50
   211.000          1          19       92.50
   247.000          1          20       97.50

First Quartile =  158.250
Median =  176.000
Third Quartile =  192.500
Interquartile range =   34.250

Alternative Methods for Obtaining Quartiles
    Method 1    2       3       4       5       6       7       8
Pcntile
Q1     157.000 158.250 157.000 159.500 160.750 157.000 160.750 157.000
Q2     176.000 176.000 176.000 176.000 176.000 176.000 176.000 176.000
Q3     191.000 192.500 191.000 192.000 191.500 191.000 191.500 193.000
NOTES:
Method 1 is the weighted average at X[np] where n is no. of cases, p is
percentile / 100
Method 2 is the weighted average at X[(n+1)p] This is used in this program.
Method 3 is the empirical distribution function.
Method 4 is called the empirical distribution function - averaging.
Method 5 is called the empirical distribution function = Interpolation.
Method 6 is the closest observation method.
Method 7 is from the TrueBasic Statistics Graphics Toolkit.
Method 8 was used in an older Microsoft Excel version.
See the internet site http://www.xycoon.com/ for the above.
========================================================
```

Frequencies

Another way to examine data is to obtain the frequency of cases that fall within categories determined by a range of score values. To do this, click on the Frequencies option under the Descriptive menu. You will see the form shown below:

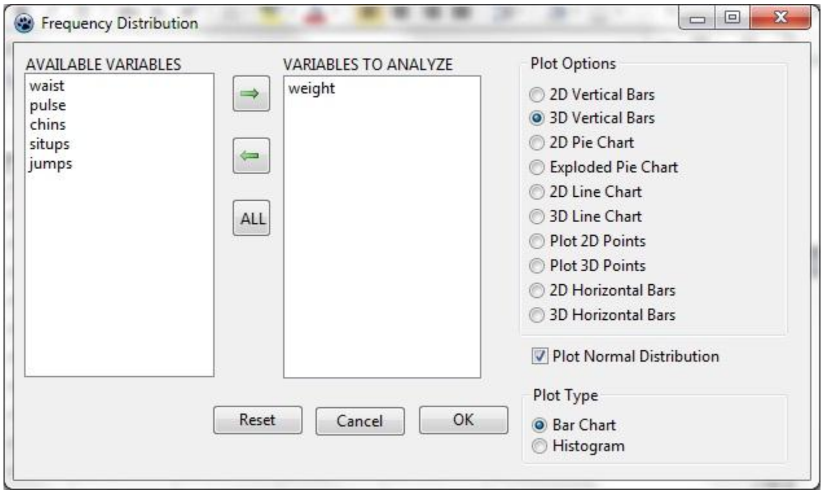


Fig. 2.2 The Frequencies Dialog

Notice that we have selected the variable "Weight" from the cansas.LAZ file. We have also elected to obtain a three dimensional, vertical bar chart of the obtained frequencies and to plot the normal distribution for corresponding frequencies behind the bar chart. Also elected was to create a new variable in the grid that contains an integer value of the frequency group. This could be useful for other graphical plots like the box plot procedure. When we click the OK button above, we first are presented with a dialog box that asks us to define the interval size and the number of intervals. One must enter an interval size that produces a number of intervals equal to or less than the number of cases. You simply click on that box and enter the new value. When you press the return key after entering a new value, you will see a change in the number of intervals. You can repeat that process until the number of intervals is acceptable. If you attempt to create more intervals than the number of cases, you will receive a warning and be returned to this dialog:

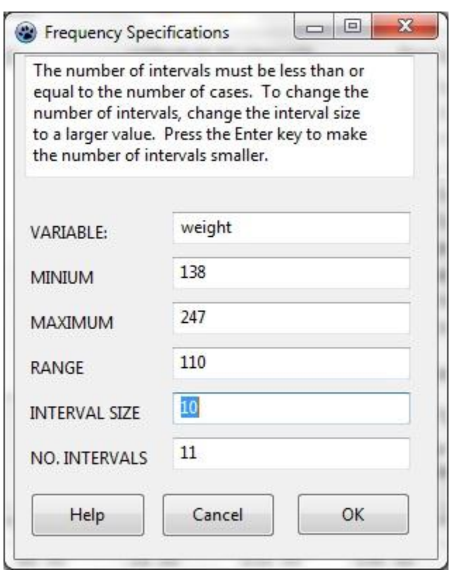


Fig. 2.3 Specifying the Interval Size and Number of Intervals for the Frequency Analysis

Notice we have changed the interval size to 10 which resulted in the number of intervals that is less than the number of cases. Clicking the OK button results in the following:

```
FREQUENCY ANALYSIS BY BILL MILLER

Frequency Analysis for weight
    FROM     TO       FREQ.    PCNT     CUM.FREQ. CUM.PCNT. %ILE RANK

  138.00  148.00       1     0.05      1.00      0.05      0.03
  148.00  158.00       4     0.20      5.00      0.25      0.15
  158.00  168.00       3     0.15      8.00      0.40      0.33
  168.00  178.00       3     0.15     11.00      0.55      0.47
  178.00  188.00       1     0.05     12.00      0.60      0.57
  188.00  198.00       5     0.25     17.00      0.85      0.72
  198.00  208.00       1     0.05     18.00      0.90      0.88
  208.00  218.00       1     0.05     19.00      0.95      0.93
  218.00  228.00       0     0.00     19.00      0.95      0.95
  228.00  238.00       0     0.00     19.00      0.95      0.95
  238.00  248.00       1     0.05     20.00      1.00      0.97

Interval ND Freq.
     1        1.16
     2        1.90
     3        2.63
     4        3.12
     5        3.14
     6        2.70
     7        1.97
     8        1.23
     9        0.65
    10        0.30
    11        0.11
    12        0.04
```

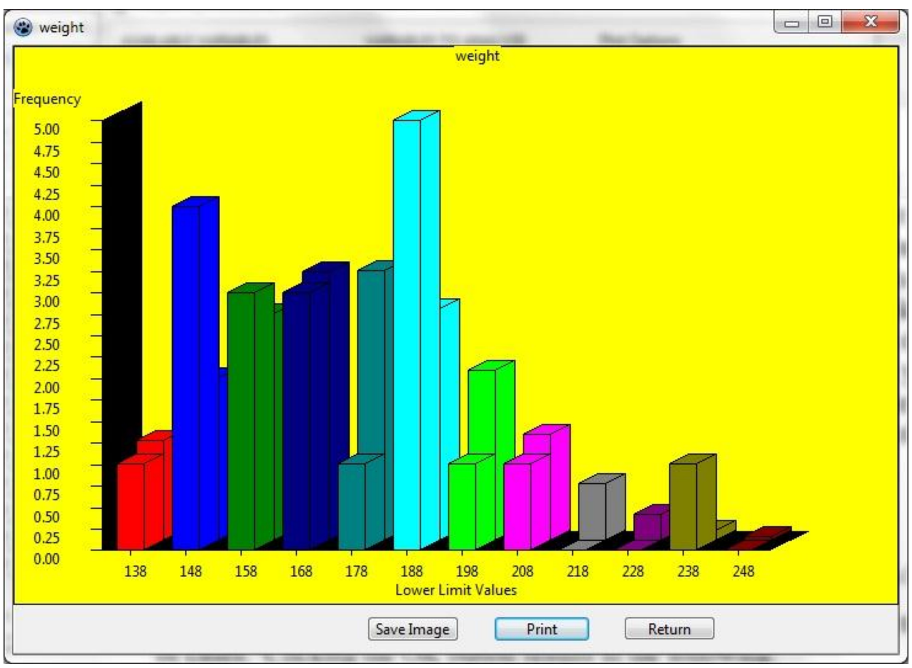


Fig. 2.4 A Plot of Frequencies in the Cansas.LAZ File

Notice that the bars in the front of the plot represent the frequency of scores in the intervals of our data while the bars behind represent frequencies expected in the normal distribution.

Cross-Tabulation

When you have entered data that represents cases classified by two or more categorical variables, it is useful to count the number of cases classified in those categories. The Cross Tabulation option of the Descriptives option gives you those results. We will use a file labeled "twoway.LAZ" to demonstrate. We have loaded the file into the grid and elected the cross tabulation option. Below are the results:

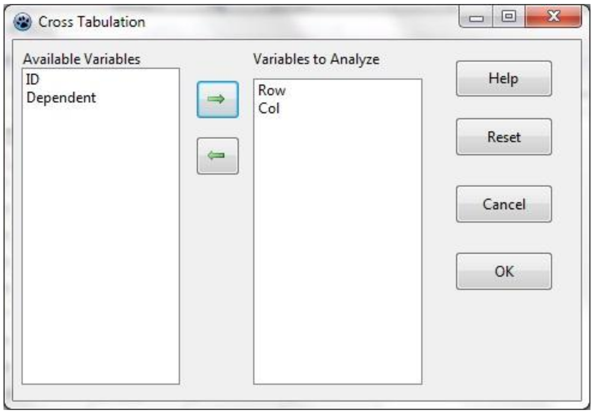


Fig. 2.5 Specification of a Cross-Tabulation

CROSSTAB RESULTS

Analyzed data is from file : C:\lazarus\Projects\LazStats\LazStatsData\twoway.LAZ

Row min.= 1, max.= 2, no. levels = 2
Col min.= 1, max.= 2, no. levels = 2

FREQUENCIES BY LEVEL:
For cell levels: Row : 1 Col: 1 Frequency = 3
For cell levels: Row : 1 Col: 2 Frequency = 3
Sum across levels = 6

For cell levels: Row : 2 Col: 1 Frequency = 3
For cell levels: Row : 2 Col: 2 Frequency = 3
Sum across levels = 6

Cell Frequencies by Levels with 12 cases.
Variables

	Col: 1	Col: 2
Block 1	3.000	3.000
Block 2	3.000	3.000

Grand sum across all categories = 32

Breakdown

A procedure related to the Cross-Tabulation procedure described above lets you analyze a continuous (floating point) variable broken down into categories of one or more categorical variables. Using the same file as above (twoway.LAZ) we will demonstrate this procedure. Below is the form and the results.

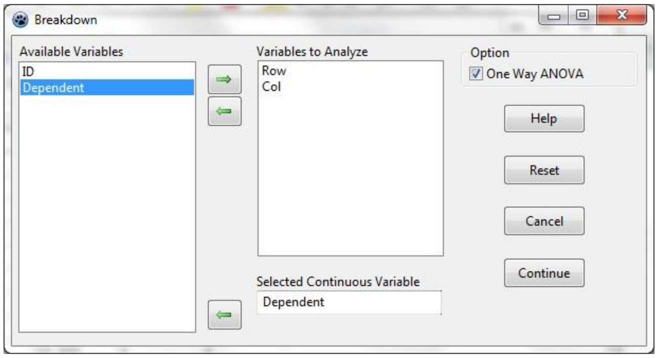


Fig. 2.6 The Breakdown Form

```
BREAKDOWN ANALYSIS PROGRAM

VARIABLE SEQUENCE FOR THE BREAKDOWN:
Row          (Variable   1) Lowest level =  1 Highest level =  2
Col          (Variable   2) Lowest level =  1 Highest level =  2

Variable levels:
Row          level =   1
Col          level =   1

Freq.     Mean     Std. Dev.
  3        3.000    1.000
Variable levels:
Row          level =   1
Col          level =   2

Freq.     Mean     Std. Dev.
  3        6.000    1.000
Number of observations accross levels =   6
Mean accross levels =    4.500
Std. Dev. accross levels =    1.871

Variable levels:
Row          level =   2
Col          level =   1

Freq.     Mean     Std. Dev.
  3       10.000    2.646
Variable levels:
Row          level =   2
Col          level =   2

Freq.     Mean     Std. Dev.
  3       12.000    2.646
Number of observations accross levels =   6
Mean accross levels =   11.000
Std. Dev. accross levels =    2.608

Grand number of observations accross all categories =  12
Overall Mean =    7.750
Overall standard deviation =    4.025

ANALYSES OF VARIANCE SUMMARY TABLES

Variable levels:
```

```
Row          level =   1
Col          level =   1

Variable levels:
Row          level =   1
Col          level =   2

SOURCE     D.F.        SS         MS         F        Prob.>F
GROUPS      1          13.50      13.50     13.500   0.0213
WITHIN      4           4.00       1.00
TOTAL       5          17.50

Variable levels:
Row          level =   2
Col          level =   1

Variable levels:
Row          level =   2
Col          level =   2

SOURCE     D.F.        SS         MS         F        Prob.>F
GROUPS      1           6.00       6.00      0.857   0.4069
WITHIN      4          28.00       7.00
TOTAL       5          34.00

ANOVA FOR ALL CELLS

SOURCE     D.F.        SS         MS         F        Prob.>F
GROUPS      3         146.25      48.75     12.188   0.0024
WITHIN      8          32.00       4.00
TOTAL      11         178.25
FINISHED
```

Normality Tests

One can test the assumption that the distribution of values in a variable are a random sample from a normally distributed population. The dialog form is shown below:

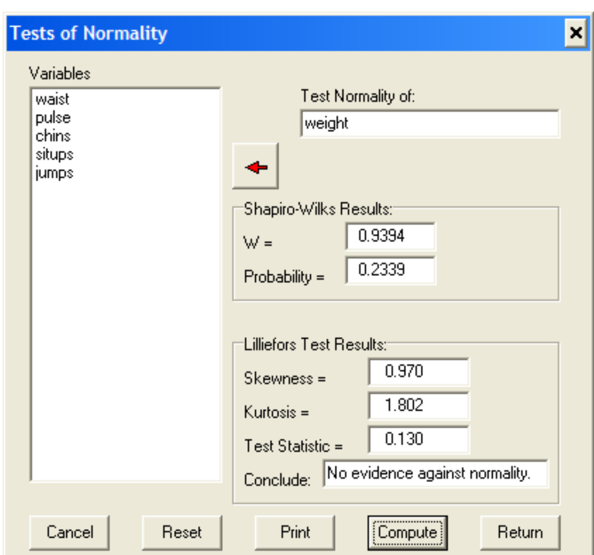


Fig. 2.7 Normality Test Dialog

In this example we have utilized the cansas.LAZ file and analyzed the weight variable. The two tests both support the assumption that weight is obtained from a normally distributed population.

X Versus Y Plot

One of the best way to examine the relationship between two variables is to plot the values of one against the other. We have selected the cansas.LAZ file and have plotted two of the variables. Shown below is the dialog form for this procedure. You can see the variables that have been selected and the options for the output that have been selected.

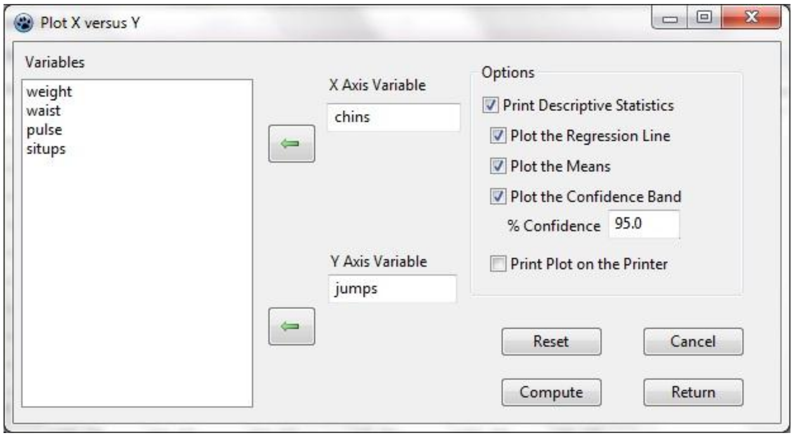


Fig. 2.8 X Versus Y Dialog

The output obtained when you click the OK button is shown below:

```
X versus Y Plot

X = chins, Y = jumps from file:
C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable       Mean   Variance  Std.Dev.
chins          9.45      27.94      5.29
jumps         70.30    2629.38     51.28
Correlation = 0.4958, Slope =      4.81, Intercept =     24.86
Standard Error of Estimate =    45.75
Number of good cases = 20
```

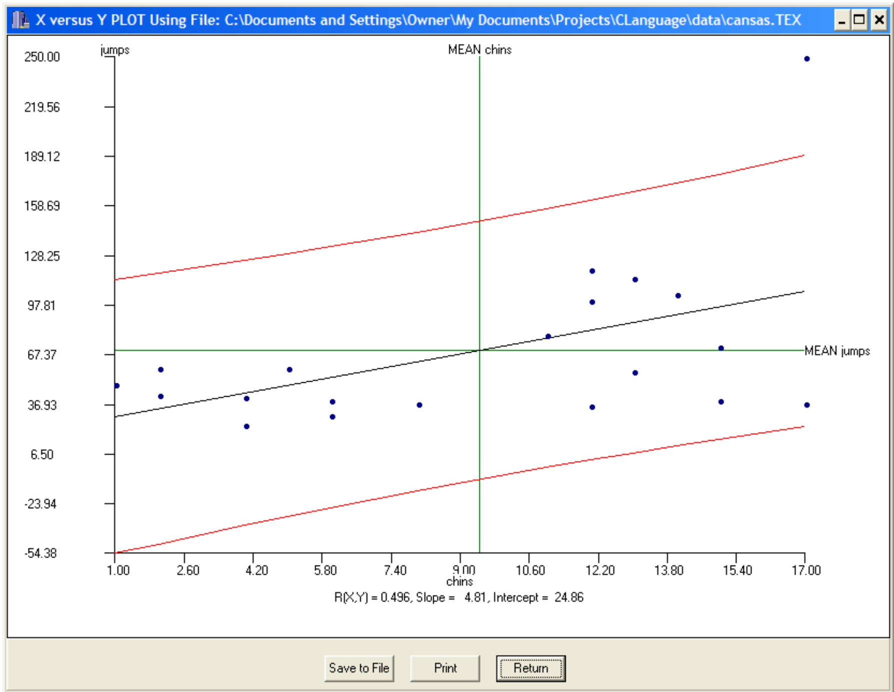


Fig. 2.9 A Plot of Two Variables

The results indicate a moderate correlation of 0.496 with considerable scatter of points. In particular, notice the "outlier" at the Y value of 17 and the jumps of 250. Elimination of that point might change the correlation quite a bit. We also notice that the pattern of points does not seem to form a symmetric oval that is expected for a bivariate-normal distribution. Notice the values below the means form a somewhat flat distribution while those above the mean for chins is more rounded. One could speculate that there might be a curvilinear relationship between these two variables. The two red curves on the border of the plots indicate the 95% confidence limits. Notice the point we mentioned lies quit a bit outside this interval.

Group Frequency Histograms

When data values have been classified as members of various groups, one can obtain a plot of the frequency of cases in each group. The frequency variable should be defined as an integer variable, typically with values from 1 to the highest group number. We have selected the file chisqr.LAZ as an example in which cases have been classified into both rows and columns. In our example we have chosen to plot the frequency of cases in the various columns and have chose the three dimensional vertical plot.

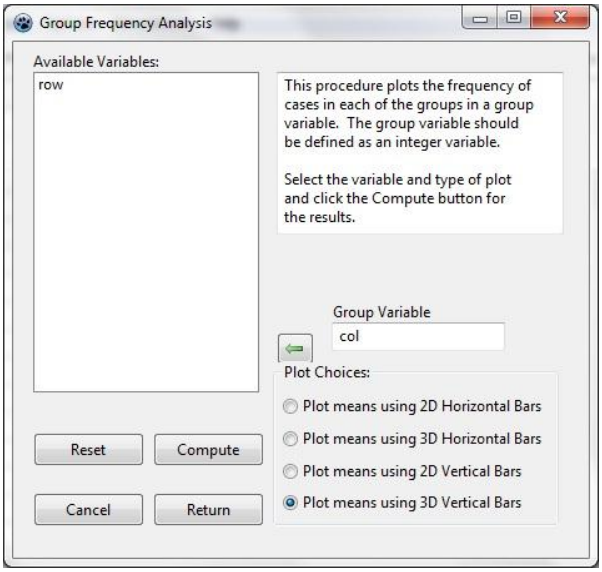


Fig. 2.10 Specification Dialog for a Frequency Analysis

The plot obtained is:

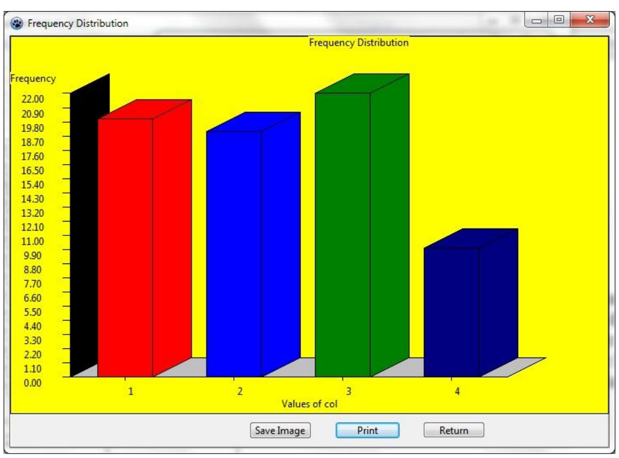


Fig. 2.11 A Sample Group Frequency Plot

Repeated Measures Bubble Plot

Teachers, physicians, economists and other professionals often collect the same measure repeatedly over time for various classes of subjects. One of the ways to examine trends in this data is to plot these repeated values with bubbles that are colored for the groups. In our example we are going to use some school data that shows achievement of students as a function of both the year the data were collected and the ratio of teachers to students. Our specifications are shown in the following dialog:

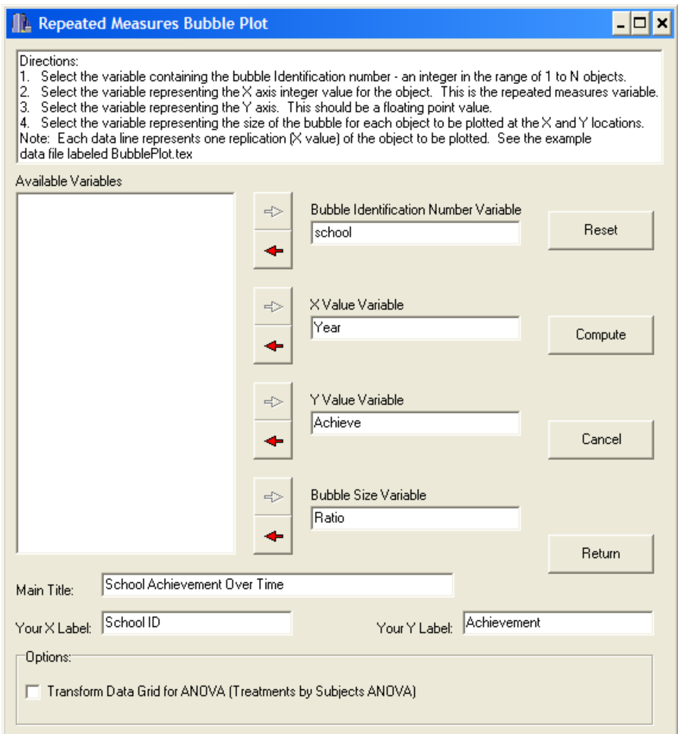


Fig. 2.12 Repeated Measures Bubble Plot Dialog

When the Compute button is clicked, the following plot is obtained:

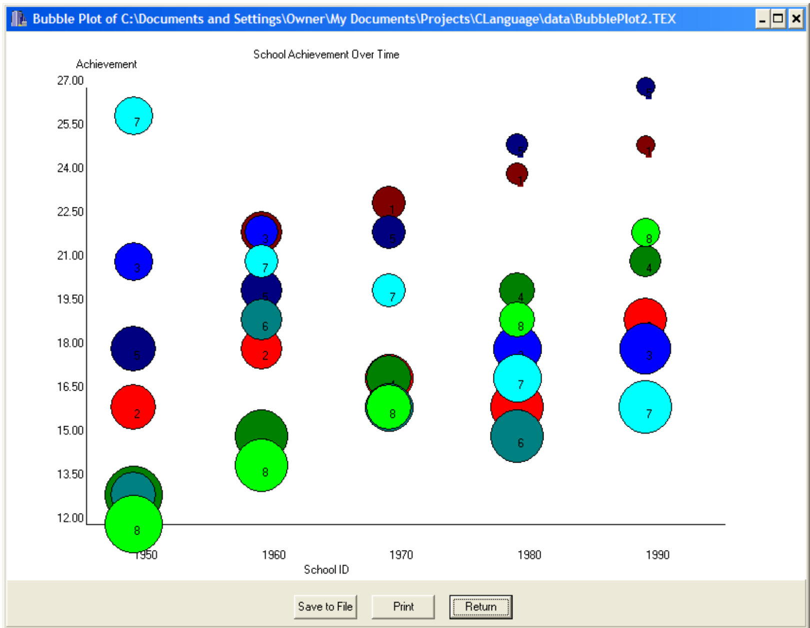


Fig. 2.13 Bubble Plot of School Achievement

Notice in this plot that as the number of students to teacher ratio increases, the acievement goes down (group 7 as an example.) Also notice the increase in achievement as the ratio decreases as demonstrated by group 8. One would most likely want to obtain the correlation between the ratio and achievement across all the years! Additional output obtained is:

```
MEANS FOR Y AND SIZE VARIABLES

Grand Mean for Y =    18.925
Grand Mean for Size =    23.125

REPLICATION MEAN Y VALUES (ACROSS OBJECTS)
Replication      1 Mean =    17.125
Replication      2 Mean =    18.875
Replication      3 Mean =    18.875
Replication      4 Mean =    19.250
Replication      5 Mean =    20.500

REPLICATION MEAN SIZE VALUES (ACROSS OBJECTS}
Replication      1 Mean =    25.500
Replication      2 Mean =    23.500
Replication      3 Mean =    22.750
Replication      4 Mean =    22.500
Replication      5 Mean =    21.375

MEAN Y VALUES FOR EACH BUBBLE (OBJECT)
Object      1 Mean =    22.400
Object      2 Mean =    17.200
Object      3 Mean =    19.800
Object      4 Mean =    17.200
Object      5 Mean =    22.400
Object      6 Mean =    15.800
Object      7 Mean =    20.000
Object      8 Mean =    16.600

MEAN SIZE VALUES FOR EACH BUBBLE (OBJECT)
```

```
Object     1 Mean =   19.400
Object     2 Mean =   25.200
Object     3 Mean =   23.000
Object     4 Mean =   24.600
Object     5 Mean =   19.400
Object     6 Mean =   25.800
Object     7 Mean =   23.200
Object     8 Mean =   24.400
```

We have plotted the ratio of student to teachers against achievement and obtained the following:

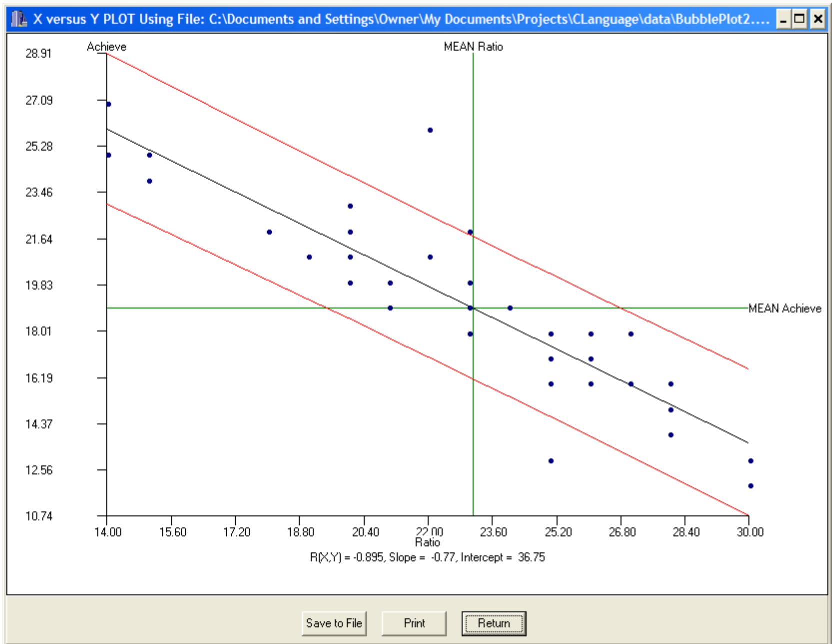


Fig. 2.14 Plot of Teacher-Student Ratio to Achievement

The above plot verifies our bubble plot which suggested a high degree of relationship between these two variables. In effect, the bubble plot is a way of viewing three dimensions of your data. In the above example we viewed the relationship among achievement (the Y axis), year (the X axis), and student to teacher ratio (the bubble size) for a number of schools (the bubbles.) You may also want to consider the three dimensional plot procedure which lets you rotate your data around the X, Y or Z axis.

Comparisons With Theoretical Distributions

LazStats lets you view the distribution of your data against a theoretical distribution in several ways. This procedure lets you plot the cumulative distribution of your data values and show the theoretical cumulative distribution of a theoretical curve. In addition, you can also plot the frequency distribution of your values against the theoretical frequency distribution. A variety of theoretical distributions are available for comparison. We will demonstrate the use of this procedure to plot the same data used previously, that is, the weight variable from the cansas.LAZ file. Show below is the dialog form:

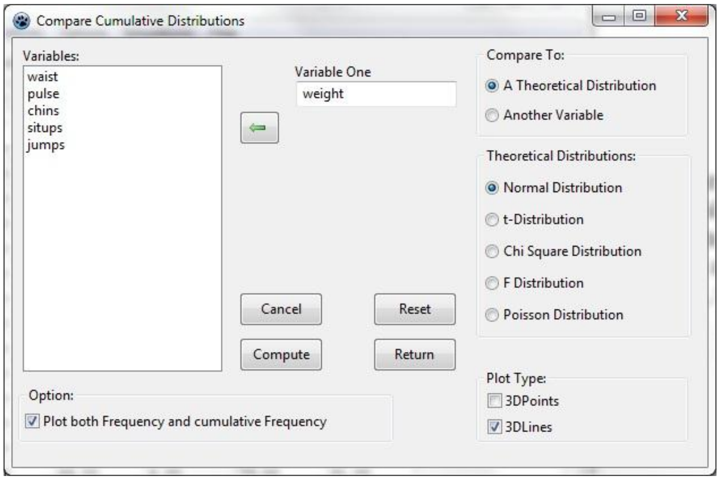


Fig. 2.15 Comparison of Cumulative Distributions

The results are:

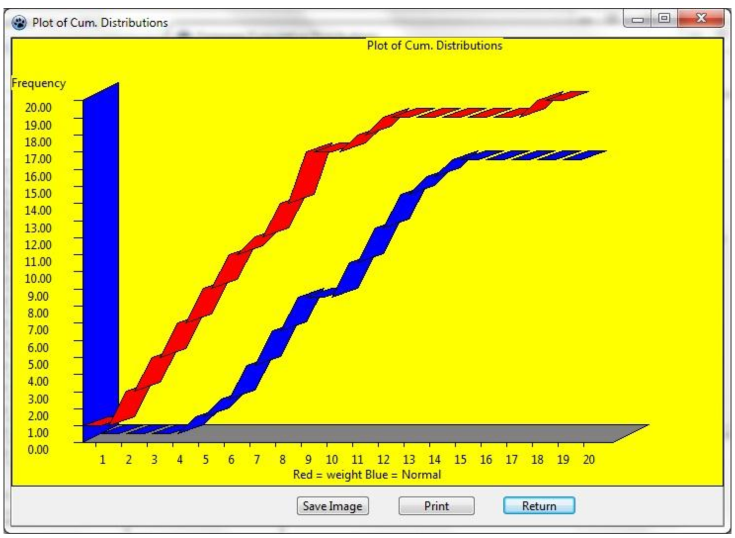


Fig. 2.16 Cumulative Normal vs. Cumulative Observed Values

Notice that the observed data seem to follow the cumulative distribution of the normal curve fairly well.

The printout for the above analysis is:

Distribution comparison by Bill Miller

weight X1 Value	weight Frequency	weight Cum. Freq.	Normal X2 Value	Normal Frequency	Normal Cum. Freq.
138.000	1	1.000	-3.000	0	0.000
144.000	0	1.000	-2.667	0	0.000
150.000	2	3.000	-2.333	0	0.000
156.000	2	5.000	-2.000	1	1.000
162.000	3	8.000	-1.667	1	2.000
168.000	1	9.000	-1.333	1	3.000
174.000	2	11.000	-1.000	2	5.000
180.000	1	12.000	-0.667	2	7.000
186.000	3	15.000	-0.333	3	10.000
192.000	2	17.000	-0.000	3	13.000
198.000	1	18.000	0.333	2	15.000
204.000	0	18.000	0.667	2	17.000
210.000	1	19.000	1.000	1	18.000
216.000	0	19.000	1.333	1	19.000
222.000	0	19.000	1.667	1	20.000

```
  228.000           0       19.000       2.000           0       20.000
  234.000           0       19.000       2.333           0       20.000
  240.000           0       19.000       2.667           0       20.000
Kolmogorov Probability = 0.765763173908239, Max Dist = 0.222222222222222
```

Three Dimensional Rotation

One gains an appreciation for the relationship among two or three variables if one can view a plot of points for three variables in a 3 dimension space. It helps even more if one can rotate those points about each of the three axis. To demonstrate we have elected three variables from the cansas.LAZ file. Show below is the dialog and plot:

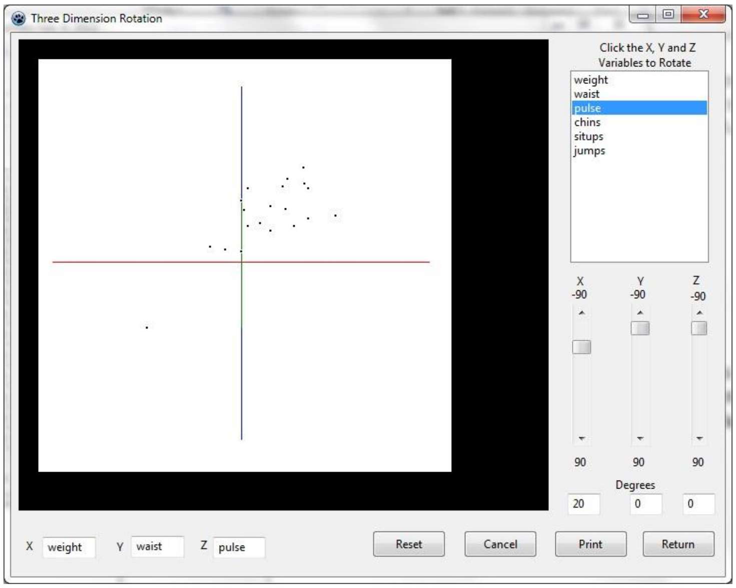


Fig. 2.17 Scatter Plot of Values for Three Variables

You can place the mouse on one of the three "scroll" bar buttons (squares in the slider portion) and drag the button down while holding down the left mouse button. This will let you see more clearly the relationships among the three variables. To demonstrate, we have rotated the Y axis to 45 and the Z axis to nearly 0 degrees to examine the relationship among the variables (weight, waist and chins.)

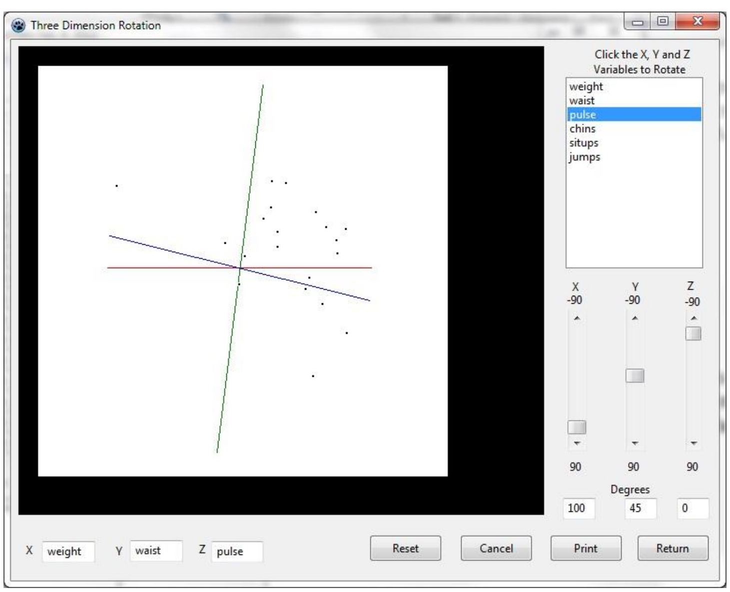


Fig. 2.18 Rotated Variables to Examine Relationship Between Two Variables

Essentially, you can rotate the points around any one of the three axis until one of the axis is hidden. This lets you see the points projected for just two of the variables at a time.

Box Plots

Box plots are a way of visually inspecting the distribution of scores within various categories. As an example, we will use a file labeled anova2.LAZ which contains data for an analysis of covariance with row, column, slice, X, covar1 and covar2 variables. We have selected to do a box plot of the X variable (the dependent variable) for the three slice categories. Shown below is the dialog box for our analysis.

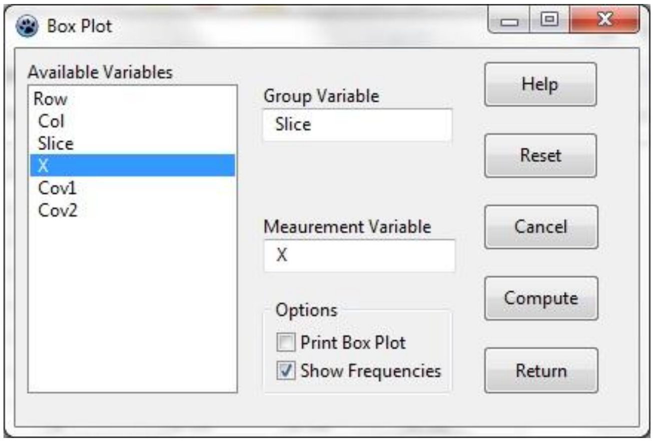


Fig. 2.19 Box Plot Dialog

Since we have elected the option of showing the frequencies within each category, we first obtain the following output:

```
Box Plot of Groups


Results for group 1, mean =    3.500
Centile        Value
Ten            1.100
Twenty five    2.000
Median         3.500
Seventy five   5.000
Ninety         5.900
Score Range     Frequency Cum.Freq. Percentile Rank
______________ _________ _________ _______________
  0.50 -   1.50   2.00      2.00       8.33
  1.50 -   2.50   2.00      4.00      25.00
  2.50 -   3.50   2.00      6.00      41.67
  3.50 -   4.50   2.00      8.00      58.33
  4.50 -   5.50   2.00     10.00      75.00
  5.50 -   6.50   2.00     12.00      91.67
  6.50 -   7.50   0.00     12.00     100.00
  7.50 -   8.50   0.00     12.00     100.00
  8.50 -   9.50   0.00     12.00     100.00
  9.50 -  10.50   0.00     12.00     100.00
 10.50 -  11.50   0.00     12.00     100.00

Results for group 2, mean =    4.500
Centile        Value
Ten            2.600
Twenty five    3.500
Median         4.500
Seventy five   5.500
```

Ninety 6.400

Score Range	Frequency	Cum.Freq.	Percentile Rank
0.50 - 1.50	0.00	0.00	0.00
1.50 - 2.50	1.00	1.00	4.17
2.50 - 3.50	2.00	3.00	16.67
3.50 - 4.50	3.00	6.00	37.50
4.50 - 5.50	3.00	9.00	62.50
5.50 - 6.50	2.00	11.00	83.33
6.50 - 7.50	1.00	12.00	95.83
7.50 - 8.50	0.00	12.00	100.00
8.50 - 9.50	0.00	12.00	100.00
9.50 - 10.50	0.00	12.00	100.00
10.50 - 11.50	0.00	12.00	100.00

Results for group 3, mean = 4.250

Centile	Value
Ten	1.600
Twenty five	2.500
Median	3.500
Seventy five	6.500
Ninety	8.300

Score Range	Frequency	Cum.Freq.	Percentile Rank
0.50 - 1.50	1.00	1.00	4.17
1.50 - 2.50	2.00	3.00	16.67
2.50 - 3.50	3.00	6.00	37.50
3.50 - 4.50	2.00	8.00	58.33
4.50 - 5.50	1.00	9.00	70.83
5.50 - 6.50	0.00	9.00	75.00
6.50 - 7.50	1.00	10.00	79.17
7.50 - 8.50	1.00	11.00	87.50
8.50 - 9.50	1.00	12.00	95.83
9.50 - 10.50	0.00	12.00	100.00
10.50 - 11.50	0.00	12.00	100.00

You can see that the procedure has obtained the centiles and percentiles for the scores in each category of our slice variable. The plot for our data is shown next:

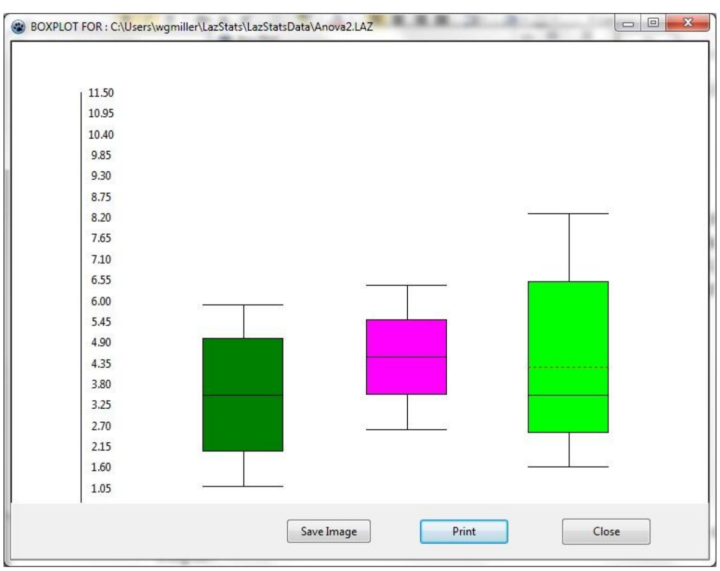


Fig. 2.20 Box Plot of the Slice Variable

The "whiskers" for each box represent the 10th and 90th percentiles. The shaded box itself represents the scores within the interquartile range. The mean and the median (50th percentile) are also plotted. In the above plot one can see that there is skewed data in the third group. The mean and median are visibly separate. The mean is the dotted line and the median is the solid line.

Plot X Versus Multiple Y Values

One often has multiple dependent measures where the measures are on a common scale of measurement or have been transformed to z scores. It is helpful to visually plot these multiple variables against an X variable common to these measures. As an example, we will use a file labeled SchoolData.LAZ. We will examine the relationship between teacher salaries and student achievement on the Scholastic Aptitude Verbal and Math scores (N = 135.) The dialog form is shown below:

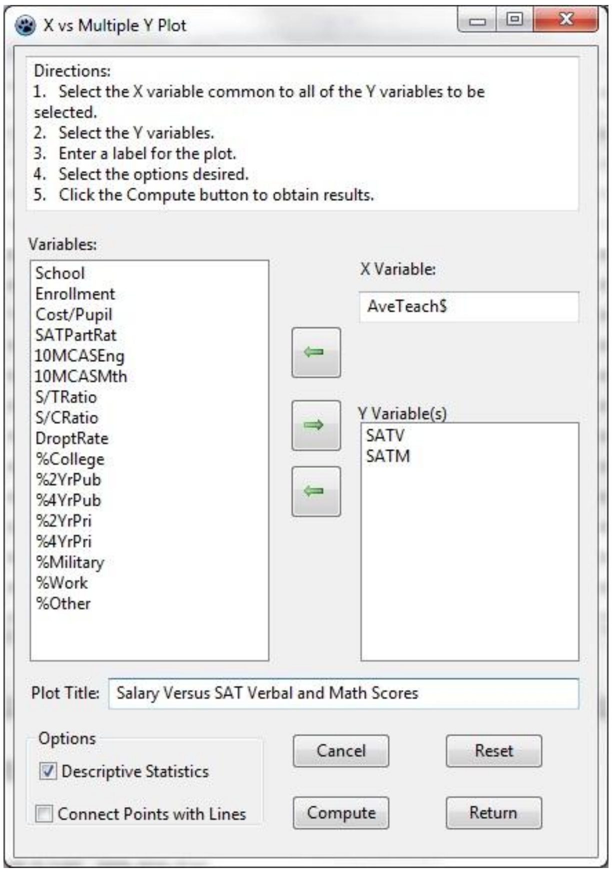


Fig. 2.21 Plot X Versus Multiple Y Dialog

Since we chose the option to show related statistics, we first obtain:

```
X VERSUS MULTIPLE Y VALUES PLOT


CORRELATION MATRIX


                 Correlations
                     SATV          SATM   AveTeach$
SATV                 1.000         0.936       0.284
SATM                 0.936         1.000       0.353
AveTeach$            0.284         0.353       1.000

Means
```

```
Variables         SATV         SATM    AveTeach$
               512.637      518.252    46963.230



Standard Deviations

Variables         SATV         SATM    AveTeach$
                41.832       44.256     4468.546


No. of valid cases = 135
```

Next we get the plot:

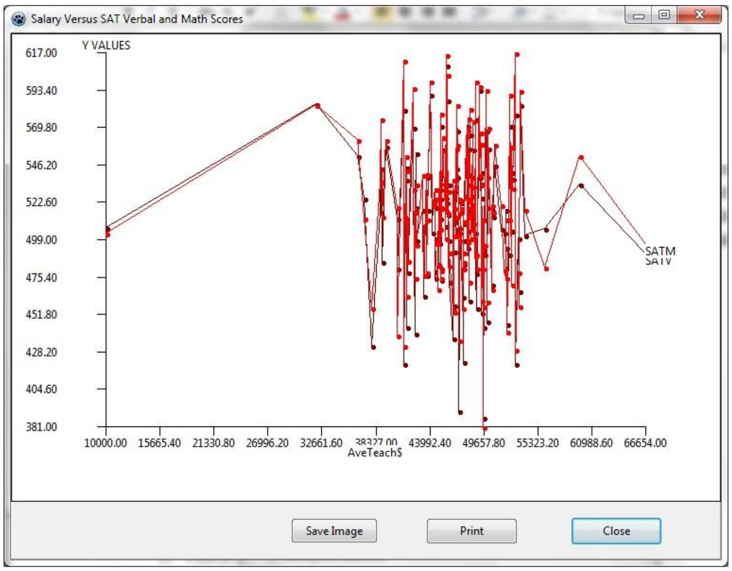


Fig. 2.22 Teacher Salaries Versus SAT Achievement

We notice several things. First we notice how closely related the two SAT scores are. Secondly, we notice a trend for higher scores as teacher salaries increase. Of course, a number of explanations could be explored to understand these relationships.

Stem and Leaf Plot

The stem and leaf plot is one of the earlier ways to graphically represent a distribution of scores for a variable. It essentially reduces the data to the two most significant digits of each value, creates a “stem” for the first (leftmost) digit and “leaves” for the second digit. If there are a large number of “leaves” for a given stem, the representation of each leaf digit may have a “depth” of more than 1 value. This prevents the plot of the individual leaf values from spilling over the right edge of your output form. The stem and leaf does give a quick view of the distribution of many variables. The example we will use is from the SchoolData.LAZ file which contains 135 cases. We will create stem and leaf plots for three of the variables in this file. The dialog form is shown below:

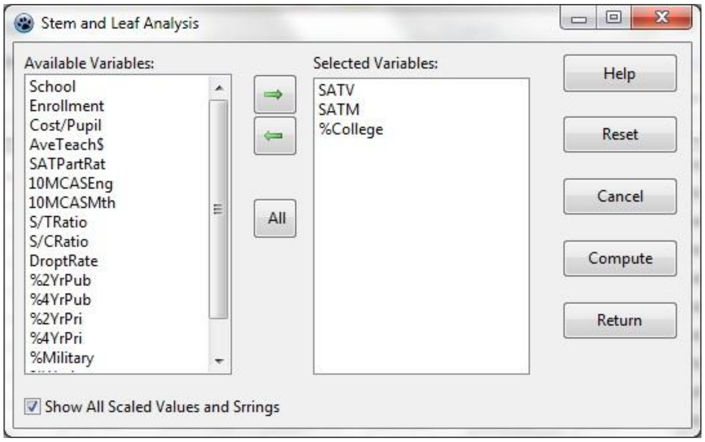


Fig. 2.23 Stem and Leaf Plot Dialog

When we click the Compute button, we obtain:

```
STEM AND LEAF PLOTS
Stem and Leaf Plot for variable: SATV

Frequency  Stem & Leaf
     2        3    89
    10        4    223444
    31        4    5566777788899999
    68        5    0000000001111111222222233333334444
    23        5    555566777889
     1        6    0

Stem width = 100.00, max. leaf depth =  2
Min. value =  387.000, Max. value =  609.000
No. of good cases = 135

Stem and Leaf Plot for variable: SATM

Frequency  Stem & Leaf
     1        3    8
     5        4    334
    40        4    555666777788888899999
    57        5    000011111111112222223333333444
    28        5    5556666777889999
     4        6    011

Stem width = 100.00, max. leaf depth =  2
Min. value =  381.000, Max. value =  617.000
No. of good cases = 135

Stem and Leaf Plot for variable: %College

Frequency  Stem & Leaf
     1        5    9
     2        6    34
     5        6    58899
    12        7    122223334444
    25        7    5555666677788888888999999
    29        8    00001111111112222233333344444
    30        8    555555566677777778888889999999
    27        9    000000011111222333444444444
```

```
     3       9    556
     1      10    0

Stem width = 10.00, max. leaf depth =  1
Min. value =   59.000, Max. value =  100.000
No. of good cases = 135
```

If we examine this last variable, we note that the stem width is 10. Now look at the top stem (5) and the leaf value 9. These are the two leftmost digits. We multiply the stem by the stem width to obtain the value 50 and then replace the second digit behind the first with the leaf value to obtain 59.. Now examine the previous plot for the SATM variable. The stem width is 100 so the first values counted are those with digits of 380. This we get by multiplying the stem width of 100 times the stem of 3 and entering the second digit of 8 behind the 3. We also note that the leaf depth is 1 in the last plot but is 2 in the previous plot. This indicates that each leaf digit in the last plot represents one value while in the previous plot each leaf represents one or two values. You might also note that the stems are "broken" into a lower half and upper half. That is, if the second digit is 0 to 4 it is plotted in the lower half of the stem value and if 5 to 9 it is plotted in the upper half of the digits for that stem.

Multiple Group X Versus Y Plot

When you have obtained data on multiple groups that includes variables possibly related, you have several choices for viewing the data graphically. One would be to plot the two variables (e.g. X and Y) against each other in a traditional X vs. Y scatter plot. This would be repeated by first selecting one group at a time. Another option would be to concurrently plot X vs. Y for all of the groups. This procedure provides this last alternative. Our example uses the BubblePlot2.LAZ file. There are eight schools that have been sampled and we wish to plot the Student to Teacher ratio (our X variable) against the Achievement variable (our Y variable.) The dialog for specifying this analysis is shown below:

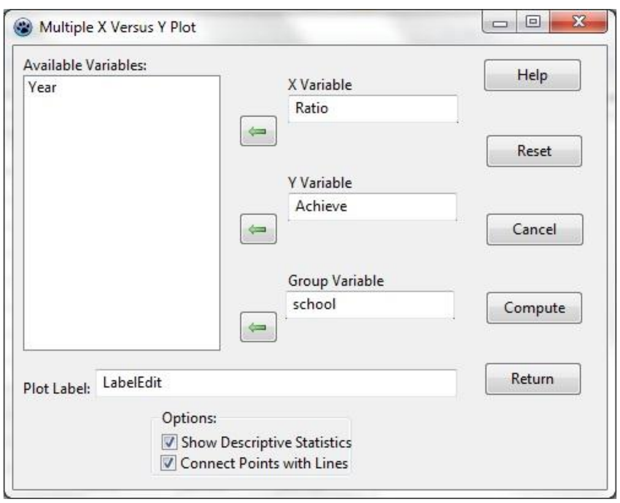


Fig. 2.24 The Multiple Group X vs. Y Plot Dialog

When we click the OK button we obtain:

```
X VERSUS Y FOR GROUPS PLOT

VARIABLE  MEAN    STANDARED DEVIATION
   X         23.125    4.268
   Y         18.925    3.675
```

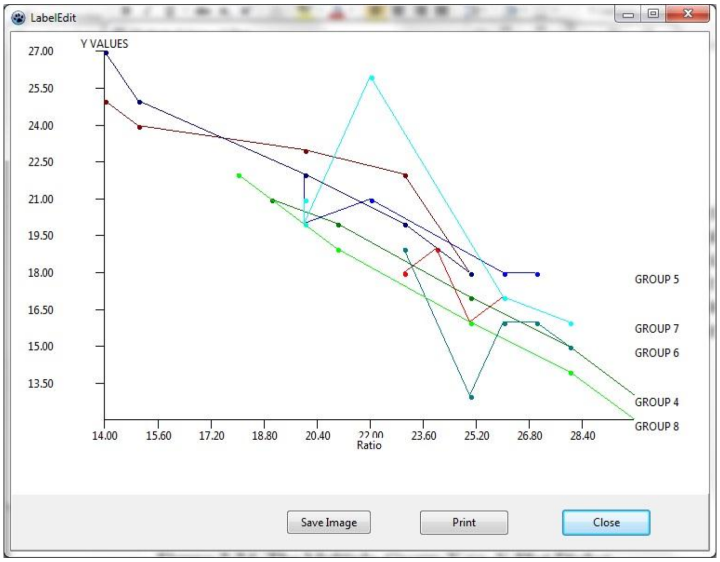


Fig. 2.25 An X vs. Y Plot for Multiple Groups

We note the common relationship among all groups that as the Student to Teacher ratio increases, the achievement of students in the schools decreases. The trend is stronger for some schools than others and this suggests we may want to complete a further analysis such as a discriminant function analysis to determine whether or not the school differences are significant.

Resistant Line

Tukey (1970, Chapter 10) proposed the three point resistant line as an data analysis tool for quickly fitting a straight line to bivariate data (x and y paired data.) The data are divided into three groups of approximately equal size and sorted on the x variable.   The median points of the upper and lower groups are fitted to the middle group to form two slope lines. The resulting slope line is resistant to the effects of extreme scores of either x or y values and provides a quick exploratory tool for investigating the linearity of the data. The ratio of the two slope lines from the upper and lower group medians to the middle group median provides a quick estimate of the linearity which should be approximately 1.0 for linearity. Our example uses the “Cansas.TEX” file. The dialogue for the analysis appears as:

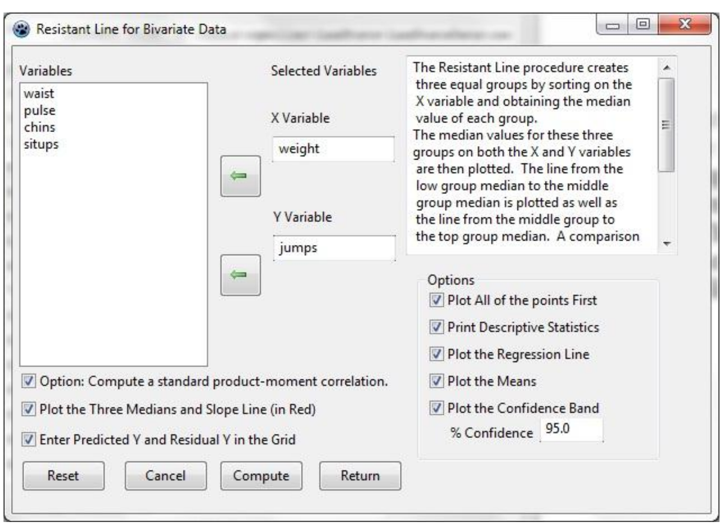


Fig. 2.26 Form for Resistant Line

The results obtained are:
Original X versus Y Plot Data

X = weight, Y = jumps from file: C:\Users\wgmiller\LazStats\LazStatsData\cansas.LAZ

```
Variable     Mean   Variance  Std.Dev.
weight     178.60    609.62     24.69
jumps       70.30   2629.38     51.28
Correlation = -0.2263, Slope =    -0.47, Intercept =   154.24
Standard Error of Estimate =    51.32
Number of good cases = 20

Group    X Median     Y Median     Size
  1      155.000      155.000      6
  2      176.000      34.000      8
  3      197.500      36.500      6

Half Slopes =      -5.762 and       0.116
Slope =     -2.788
Ratio of half slopes =      -0.020
Equation: y =     -2.788 * X + ( -566.361)
```

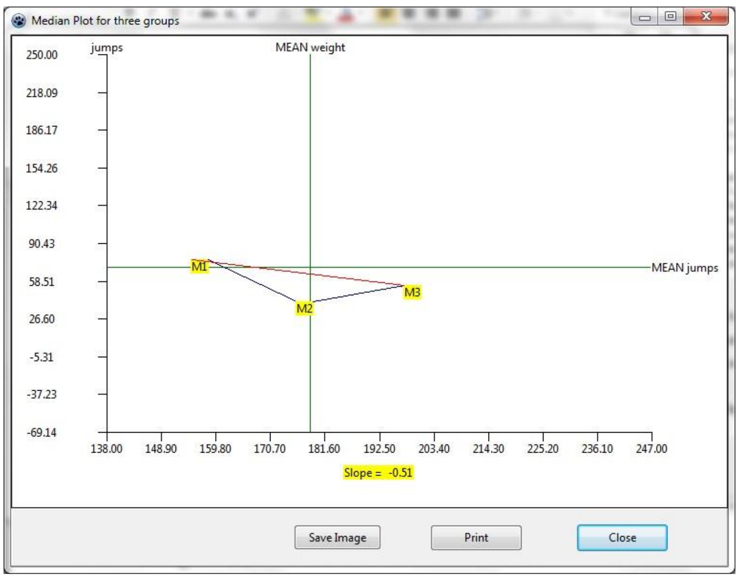


Fig. 2.27 Median Plot for Resistant Line

Notice that the estimated slope of the resistant line is slightly different than that obtained from the traditional correlation analysis.

Compare Distributions

It may be desirable to compare the distribution of two continuous variables. As an example, we have loaded a file labeled anova2.LAZ and transformed the dependent variable x to ranks. We then wish to compare the original value x with the rankings on x. The dialog for the analysis is shown below with the output after that.

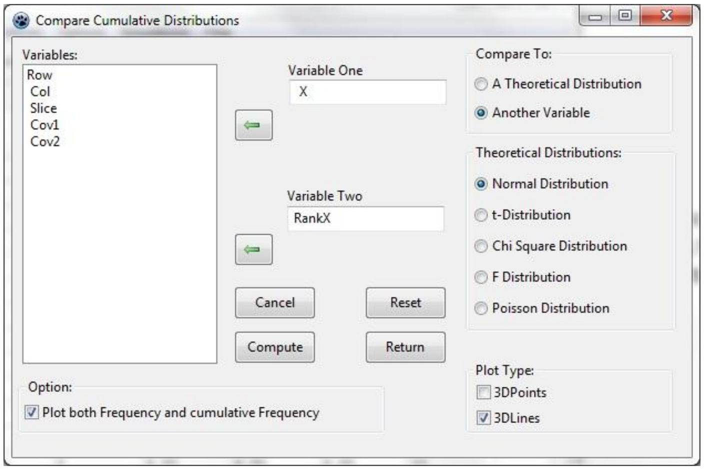


Fig. 2.28 Form for Comparing Distributions

Distribution comparison by Bill Miller

X1 Value	Frequency	Cum. Freq.	RankX X2 Value	RankX Frequency	RankX Cum. Freq.
1.000	3	3.000	2.000	3	3.000
1.450	0	3.000	3.750	0	3.000
1.900	5	8.000	5.500	5	8.000
2.350	0	8.000	7.250	0	8.000
2.800	7	15.000	9.000	0	8.000
3.250	0	15.000	10.750	7	15.000
3.700	7	22.000	12.500	0	15.000
4.150	0	22.000	14.250	0	15.000
4.600	6	28.000	16.000	0	15.000
5.050	0	28.000	17.750	7	22.000
5.500	0	28.000	19.500	0	22.000
5.950	4	32.000	21.250	0	22.000
6.400	0	32.000	23.000	0	22.000
6.850	2	34.000	24.750	6	28.000
7.300	0	34.000	26.500	0	28.000
7.750	1	35.000	28.250	0	28.000
8.200	0	35.000	30.000	4	32.000
8.650	1	36.000	31.750	0	32.000
9.100	0	36.000	33.500	3	35.000
9.550	0	36.000	35.250	1	36.000

Kolmogorov Probability = 0.55955971019521, Max Dist = 0.25

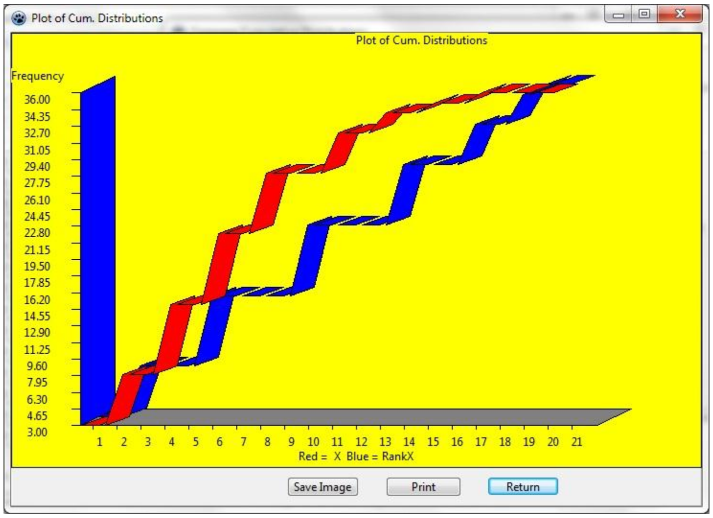


Fig. 2.29 Plot of Cumulative Distributions

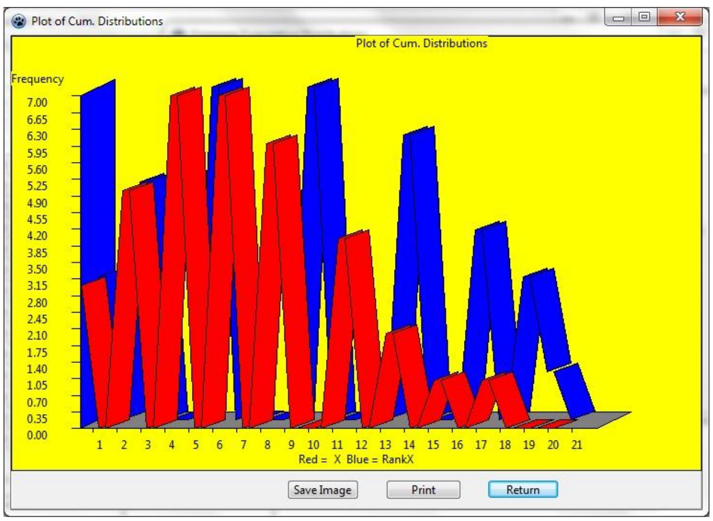


Fig. 2.30 Frequency Plot of Two Distributions

Data Smoothing

Data obtained may often contain “noise” that masks the major variations in a variable. This noise may be reduced by “smoothing” the data. Several techniques have been developed for this task. The one included in this package involves the averaging of three contiguous values at a time and replacing the lead value with the average. This is repeated across the values of the variable (with the exception of the first and last values.) To demonstrate, we have loaded a file labeled boltsize.LAZ and will smooth the VAR1 variable. The dialog is shown below and a comparison plot of the original and smoothed data is also shown.

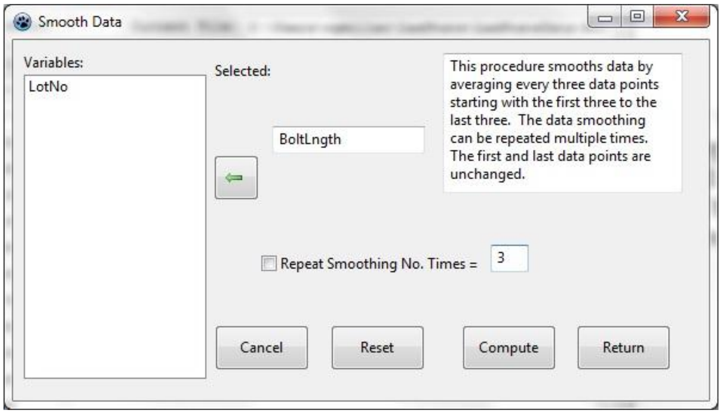


Fig. 2.31 Form for Smoothing Data

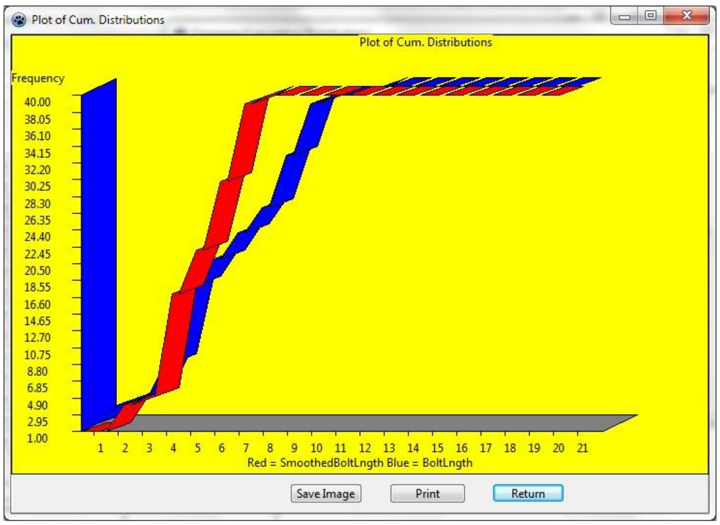


Fig. 2.32 Smoothed Data Cumulative Distributions

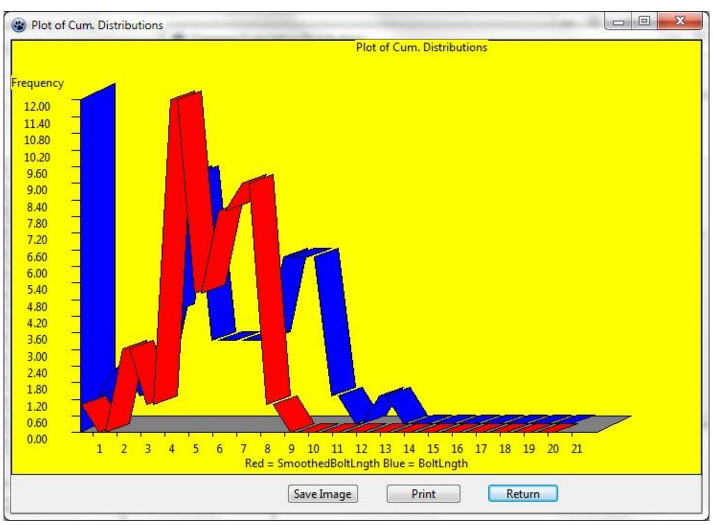


Fig. 2.33 Smoothed Frequency Distributions

Chapter 3. Comparisons

Testing Hypotheses for Differences Between or Among Means

The Nature of Scientific Investigation.

People have been trying to understand the things they observe for as long as history has been recorded. Understanding observed phenomenon implies an ability to describe and predict the phenomenon. For example, ancient man sought to understand the relationship between the sun and the earth. When man is able to predict an occurrence or change in something he observes, it affords him a sense of safety and control over events. Religion, astrology, mysticism and other efforts have been used to understand what we observe. The scientific procedures adopted in the last several hundred years have made a large impact on human understanding. The scientific process utilizes inductive and deductive logic and the symbols of logic, mathematics. The process involves:

(a) Making systematic observations (**description**)
(b) Stating possible relationships between or differences among objects observed (**hypotheses**)
(c) Making observations under controlled or natural occurrences of the variations of the objects hypothesized to be related or different (**experimentation**)
(d) Applying an accepted decision rule for stating the truth or falsity of the speculations (**hypothesis testing**)
(e) Verifying the relationship, if observed (**prediction**)
(f) Applying knowledge of the relationship when verified (**control**)
(g) Conceptualizing the relationship in the context of other possible relationships (**theory**).

The rules for deciding the truth or falsity of a statement utilizes the assumptions developed concerning the chance occurrence of an event (observed relationship or difference). These decision rules are particularly acceptable because the user of the rules can ascertain, with some precision, the likelihood of making an error, whichever decision is made!

As an example of this process, consider a teacher who observes characteristics of children who mark false answers true in a true-false test as different from children who mark true answers as false. Perhaps the hypothetical teacher happens to notice that the proportion of left-handed children is greater in the first group than the second. Our teacher has made a systematic observation at this point. Next, the teacher might make a **scientific statement** such as "Being left-handed increases the likelihood of responding falsely to true-false test items." Another way of making this statement however could be "The proportion of left-handed children selecting false options of true statements in a true-false test does not differ from that of right handed children beyond that expected by sampling variability alone." This latter statement may be termed a **null hypothesis** because it states an absence (null) of a difference for the groups observed. The null hypothesis is the statement generally accepted for testing because the alternatives are innumerable. For example (1) no difference exists or (2) some difference exists. The scientific statement which states the principle of interest would be difficult to test because the possible differences are innumerable. For example, "increases" in the example above is not specific enough. Included in the set of possible "increases" are 0.0001, 0.003, 0.012, 0.12, 0.4, etc. After stating the null hypothesis, our scientist-teacher would make **controlled observations**. For example, the number of "false" options chosen by left and right handed children would be observed after controlling for the total number of items missed by each group. This might be done by **matching** left handed children with right handed children on the total test scores. The teacher may also need to insure that the number of boys and girls are also matched in each group to control for the possibility that sex is the variable related to option choices rather than handedness. We could continue to list other ways to control our observations in order to rule out variables other than the hypothesized ones possibly affecting our **decision**.

Once the teacher has made the controlled observations, decision rules are used to **accept or reject** the null hypothesis. We will discover these rules involve the chances of rejecting a true null hypothesis (**Type I error**) as well as the chances of accepting a false null hypothesis (**Type II error**).

Because of the chances of making errors in applying our decision rules, results should be verified through the observation of additional samples of subjects.

Decision Risks.

Many research decisions have different losses which may be attached to outcomes of an experiment. The Fig. below summarizes the possible outcomes in testing a null hypothesis. Each outcome has a certain probability of occurrence. These probabilities (chances) of occurrence are symbolized by Greek letters in each outcome cell.

Possible Outcomes of an Experiment

		True State of Nature	
		H_o True	H_o False
Experimenter conclusion based on observed data	accept H_o	$1 - \alpha$	ß Type II error
	reject H_o	Type I Error α	1 - ß

In the above Fig. α (alpha) is the chance of obtaining a sample which leads to rejection of the null hypothesis when in the population from which the sample is drawn the null hypothesis is actually true. On the other hand, we also have the chance of drawing a sample that leads us to accept a null hypothesis when, in fact, in the population we should reject it. This latter error has ß (Beta) chances of occurring. Greek symbols have been used rather than numbers because the experimenter may control the types of error! For example, by selecting large samples, by reducing the standard deviation of the observed variable (for example by improving the precision of measurement), or by decreasing the size of the discrepancy (difference) we desire to be sensitive to, we can control both Type I and Type II error.

Typically, the chances of getting a Type I error is arbitrarily set by the researcher. For example, the value of alpha may be set to .05. Having set the value of α, the researcher can establish the sample size needed to control Type II error which is also arbitrarily chosen (e.g. ß = .2). In other cases, the experimenter is limited to the sample size available. In this case the experimenter must also determine the smallest difference or effect size (alternate hypothesis) to which he or she wishes to be sensitive.

How does a researcher decide on α, ß and a minimum discrepancy? By assessing or estimating the loss or consequences in making each type of error! For example, in testing two possible cancer treatments, consider that treatment 1 costs $1,000 while treatment 2 costs $100. Consider the null hypothesis

H_o: no difference between treatments (i.e. equally effective)

and consider the alternative

H_1: treatment 1 is more effective than treatment 2.

If we reject H_o: and thereby accept H_1: we will pay more for cancer treatment. We would probably be glad to do this if treatment 1 were, in fact, more effective. But if we have made a Type I error, our losses are 10 to 1 in dollars lost. On the other hand, consider the loss if we should accept H_0: when, in fact, H_1: is correct. In this case lives will be lost that might have been saved. What is one life worth? Most people would probably place more than

$1,000 value on a life. If so, you would probably choose a smaller ß value than for α. The size of both these values are dependent on the size of risk you are willing to take. In the above example, a ß = .001 would not be unreasonable.

Part of our decision concerning α and ß also is based on the cost for obtaining each observation. Sometimes **destructive** observation is required. For example, in testing the effectiveness of a manufacturer's military missiles, the sample drawn would be destroyed by the testing. In these cases, the cost of additional observations may be as large as the losses associated with Type I or Type II error!

Finally, the size of the discrepancy selected as "meaningful" will affect costs and error rates. For example, is an IQ difference of 5 points between persons of Group A versus Group B a "practical" difference? How much more quickly can a child of 105 IQ learn over a child of 100 IQ? The larger the difference selected, the smaller is the sample needed to be sensitive to true population differences of that size. Thus, cost of data collection may be conserved by selecting realistic differences for the alternative hypothesis. If sample size is held constant while the discrepancy is increased, the chance of a Type II error is reduced, thus reducing the chances of a loss due to this type of error. We will examine the relationships between Type I and Type II error, the discrepancy chosen for an alternative hypothesis, and the sample size and variable's standard deviation in the following sections.

Hypotheses Related to a Single Mean.

In order to illustrate the principles of hypothesis testing, we will select an example that is rather simple. Consider a hypothetical situation of the teacher who has administered a standardized achievement test in algebra to high school students completing their first course in algebra. Assume that extensive "norms" exist for the test showing that the population of previously tested students obtained a mean score equal to 50 and a standard deviation equal to 10. Further assume the teacher has 25 students in the class and that the class test mean was 55 and the standard deviation was 9. The teacher feels that his particular method of instruction is superior to those used by typical instructors and results in superior student performance. He wishes to provide evidence for his claim through use of the standardized algebra test. However, other algebra teachers in his school claim his teaching is really no better than theirs but requires half again as much time and effort. They would like to see evidence to substantiate their claim of no difference. What must our teachers do? The following steps are recommended by their school research consultant:

1. Agree among themselves how large a difference between the past population mean and the mean of the sampled population is a practical increment in algebra test performance.
2. Agree upon the size of Type I error they are willing to accept considering the consequences.
3. Because sample size is already fixed (n=25), they cannot increase it to control Type II error. They can however estimate what it will be for the alternative hypothesis that the sampled population mean does differ by a value as large or larger than that agreed upon in (2) above.
4. Use the results obtained by the classroom teacher to accept or reject the null hypothesis assuming that the sample means of the kind obtained by the teacher are normally distributed and unbiased estimates of the population mean. This is equivalent to saying we assume the teacher's class is a randomly selected sample from a population of possible students taught be the instructor's method. We also assume that the effect of the instructor is independent for each student, that is, that the students do not interact in such a way that the score of one student is somehow

dependent on the score obtained by another
student.

By assuming that sample means are normally distributed, we may use the probability distribution of the normally distributed z to test our hypothesis. Based on a theorem known as the "Central Limit Theorem", it can be demonstrated that sample means obtained from scores that are NOT normally distributed themselves DO tend to be normally distributed! The larger the sample sizes, the closer the distribution of sample means approaches the normal distribution. You may remember that our z score transformation is

$$z = \frac{X - \bar{X}}{S_X} = \frac{d}{S_X} \tag{3.1}$$

when determining an individual's z score in a sample. Now consider our possible sample means in the above experiment to be individual scores that deviates (d) from a population mean (μ) and have a standard deviation equal to

$$S_{\bar{X}} = \frac{S_X}{\sqrt{n}} \tag{3.2}$$

That is, the sample means vary inversely with the square root of the sample size. The standard deviation of sample means is also called the standard error of the mean. We can now transform our sample mean (55) into a z score where $\mu = 50$ and the standard error is $S_e = S_x / \sqrt{n} = 10 / 5 = 2$. Our result would be:

$$z_0 = \frac{\bar{X} - \mu_0}{S_e} = \frac{55 - 50}{2} = 2.5 \tag{3.3}$$

Note we have used a small zero subscript by the population mean to indicate this is the null hypothesis mean.

Before we make any inference about our teacher's student performance, let us assume that the teachers agreed among themselves to set the risk of a Type I error rather low, at .05, because of the inherent loss of greater effort and time on their part if the hypothesis is rejected (assuming they adopt the superior teaching method). Let us also assume that the teachers have agreed that a class that achieves an average mean at least 2 standard deviations of the sample means above the previous population mean is a realistic or practical increment in algebra learning. This means that the teachers want a difference of at least 4 points from the mean of 50 since the standard error of the means is 2.

Now examine the Fig. below. In this Fig. the distribution of sample means is shown (since the statistic of interest is the sample mean.) A small caret (^) is shown at the scale point where our specific sample statistic (the mean) falls in the theoretical distribution that has a mean of 50 and standard error of 2. Also shown, by shading is the area corresponding to the extreme .05 area of the distribution.

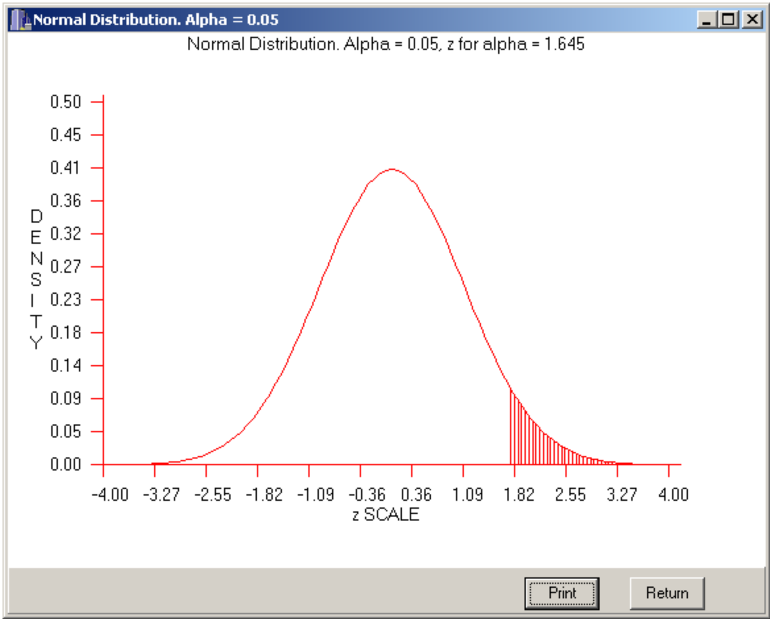


Fig. 3.1 Distribution of Sample Means

Examination of the previous Fig. indicates that the sample mean obtained deviates from the hypothesized mean by a considerable amount (5 points). If we were obtaining samples from a population in which the mean was 50 and the standard error of the means was 2, we would expect to obtain a sample this deviant only .006 of the time! That is, only .006 of normally distributed z scores are as large or larger than the z = 2.5 that we obtained! Because our sample mean is *SO* deviant for the hypothesized population, we reject the hypothesized population mean and instead accept the alternative that the population from which we did sample has a mean greater than 50. If our statistic had not exceeded the z score corresponding to our Type I error rate, we would have accepted the null hypothesis. Using a table of the normally distributed z score you can observe that the critical value for our decision is a $z_{\square}$ = 1.645.

To summarize our example, we have thus far:

1. Stated our hypothesis. In terms of our critical z score corresponding to μ, we may write the hypothesis as

$$H_0:\ z < z_\mu$$

(3.4)

2. Stated our alternate hypothesis which is

$$H_1:\ z \geq z_\mu$$

3. Obtained sample data and found that $z > z_\mu$ which leads us to reject H_0: in favor of H_1: .

Determining Type II Error and Power of the Test

In the example described above, the teachers had agreed that a deviation as large as 2 times the standard deviation of the means would be a "practical" teaching gain. The question may be asked, "What is the probability of accepting the null hypothesis when the true population mean is, in fact, 2 standard deviations (standard error) units above the hypothesized mean?" The Fig. below illustrates the theoretical distributions for both the null hypothesis and a specific alternate hypothesis, i.e. H_1= 54.

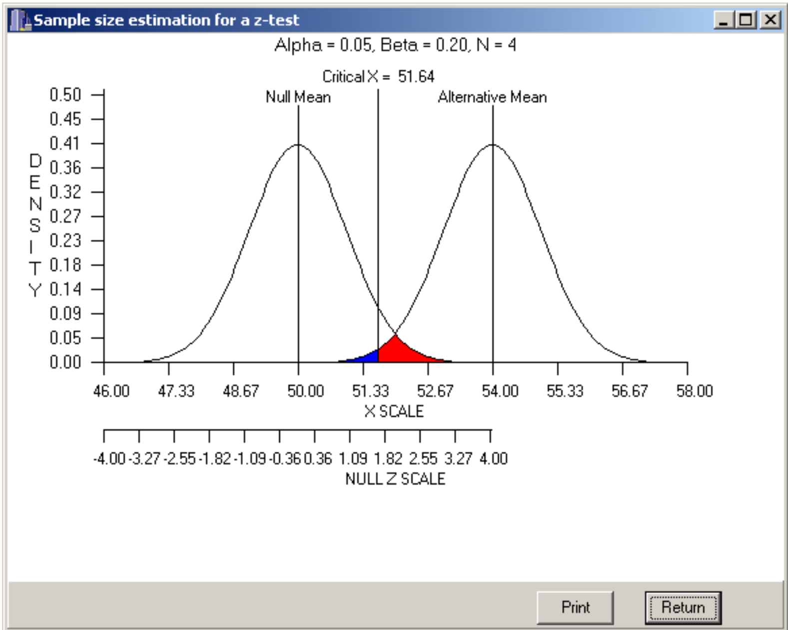


Fig. 3.2 Sample Size Estimation

The area to the left of the α value of 1.645 (frequently referred to as the region of rejection) under the null distribution (left-most curve) is the area of "acceptance" of the null hypothesis - any sample mean obtained that falls in this region would lead to acceptance of the null hypothesis. Of course, any sample mean obtained that is larger than the z = 1.645 would lead to rejection (the shaded portion of the null distribution). Now we may ask, "If we consider the alternative distribution (i.e. μ = 54), what is the z value in <u>that</u> distribution which corresponds to the z value for μ under the null distribution?" To determine this value, we will first transform the z score for alpha under the null distribution back to the raw score X to which it corresponds. Solving the z score formula for X we obtain

$$\bar{X} = z_{\mu} S_{\bar{X}} + \mu_0$$

$$\text{or} \quad \bar{X} = 1.645\,(2) + 50 = 53.29 \tag{3.5}$$

Now that we have the raw score mean for the critical value of alpha, we can calculate the corresponding z score under the alternate distribution, that is

$$z_1 = \frac{\bar{X} - \mu_1}{S_{\bar{X}}} = \frac{53.29 - 54}{2} = -.355 \tag{3.6}$$

We may now ask, "What is the probability of obtaining a unit normal z score less than or equal to -.355?" Using a table of the normal distribution or a program to obtain the cumulative probability of the z distribution we observe that the probability is ß = .359. In other words, the probability of obtaining a z score of -.355 or less is .359 under the normal distribution. We conclude then that the Type II error of our test, that is, the probability of <u>incorrectly</u> accepting the null hypothesis when, in fact, the true population mean is 54 is .359. Note that this nearly 36% chance of an error is considerably larger than the 5% chance of making the Type I error!

The sensitivity of our statistical test to detect true differences from the null hypothesized value is called the Power of our test. It is obtained simply as 1 - ß. For the situation of detecting a difference as large as 4 (two standard deviations of the sample mean) in our previous example, the power of the test was 1 - .359 = .641. We may, of course, determine the power of the test for many other alternative hypotheses. For example, we may wish to know the power of our test to be sensitive to a discrepancy as large as 6 X score units of the mean. The Fig. below illustrates the power curves for different Type I error rates and differences from the null hypothesis.

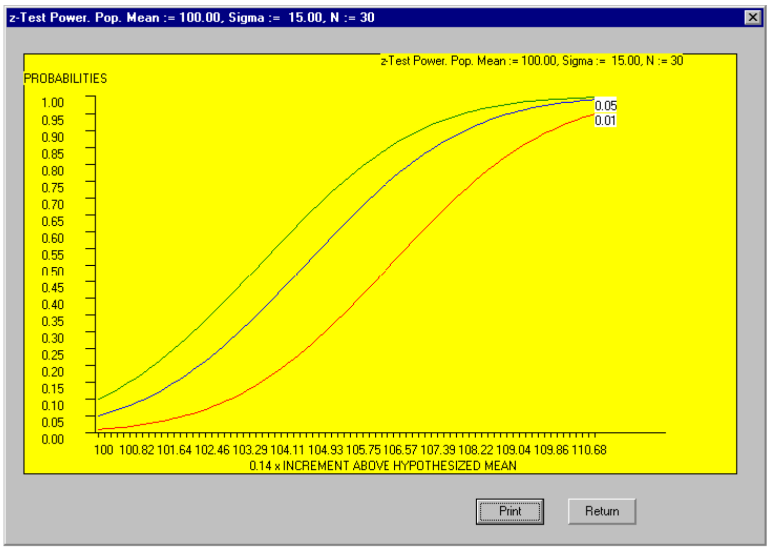


Fig. 3.3 Power Curves

Again, our procedure for obtaining the power would be

a) Obtain the raw X-score mean corresponding to the critical value of α (region of rejection) under the null hypothesis. That is

$$\bar{X} = z_\alpha S_{\bar{X}} + \mu_0 \qquad (3.7)$$

$$= 1.645\ (2) + 50 = 53.29$$

b) Obtain the z_1 score equivalent to the critical raw score for the alternate hypothesized distribution, e.g.

$$z_1 = (\bar{X} - \mu_1) / S_{\bar{X}} \qquad (3.8)$$

$$= (53.29 - 56) / 2$$

$$= -2.71 / 2$$

$$= -1.355$$

c) Determine the probability of obtaining a more extreme value than that obtained in (b) under the unit-normal distribution, e.g.

$P(z < z_1 \mid ND: \mu = 0, \sigma = 1) =$

$$P(z < -1.355 \mid ND: \mu = 0, \sigma = 1) = .0869 \quad (3.9)$$

d) Obtain the power as $1 - ß = 1.0 - .0869 = .9131$ (3.10)

One may repeat the above procedure for any number of alternative hypotheses and plot the results in a Fig. such as that shown above. The above plot was made using the LazStats option labeled "Generate Power Curves" in the Utilities menu.

As the critical difference increases, the power of the test to detect the difference increases. Minimum power is obtained when the critical difference is equal to zero. At that point power is equal to α, the Type I error rate. A different "power curve" may be constructed for every possible value of α. If larger values of α are selected, for example .20 instead of .05, then the test is more powerful for detecting true alternative distributions given the same meaningful effect size, standard deviation and sample size.

The Fig. 6 above shows the power curves for our example when selecting the following values of α: .01, .05, and .10.

Sample Size Requirements for the Test of One Mean

The translation of a raw score mean into a standard score was obtained by

$$z = \frac{\bar{X} - \mu}{S_{\bar{X}}} \quad (3.11)$$

Likewise, the above formula may be rewritten for translating a z score into the raw score mean by:

$$\bar{X} = S_{\bar{X}} z + \mu \quad (3.12)$$

Now consider the distribution of an infinite number of sample means where each mean is based on the same number of randomly selected cases. Even if the original scores are not from a normally distributed population, if the means are obtained from reasonably large samples ($N > 30$), the means will tend to be normally distributed. This phenomenon is known as the *Central Limit Theorem* and permits us to use the normal distribution model in testing a wide range of hypotheses concerning sample means.

The extreme "tails" of the distribution of sample means are sometimes referred to as "critical regions". Critical regions are defined as those areas of the distribution which are extreme, that is unlikely to occur often by chance, and which represent situations where you would reject the distribution as representing the true population should you obtain a sample in that region. The size of the region indicates the proportion of times sample values would result in rejection of the null hypothesis by chance alone - that is, result in a "Type I" error. For the situation of our last example, the full region "R" of say .05 may be split equally between both tails of the distribution, that is, .025 or R / 2 is in each tail. For normally distributed statistics a .025 extreme region corresponds to a z score of either -1.96 for the lower tail or +1.96 for the upper tail. The critical sample mean values that correspond to these regions of rejection are therefore

$$\bar{X}_c = \pm \sigma_{\bar{X}} z_{\alpha/2} + \mu_0 \quad (3.13)$$

In addition to the possibility of a critical score (X_C) being obtained by chance part of the time (α) there also exists the probability (ß) of accepting the null hypothesis when in fact the sample value is obtained from a population with a mean different from that hypothesized. Carefully examine the Fig. 3.4 below.

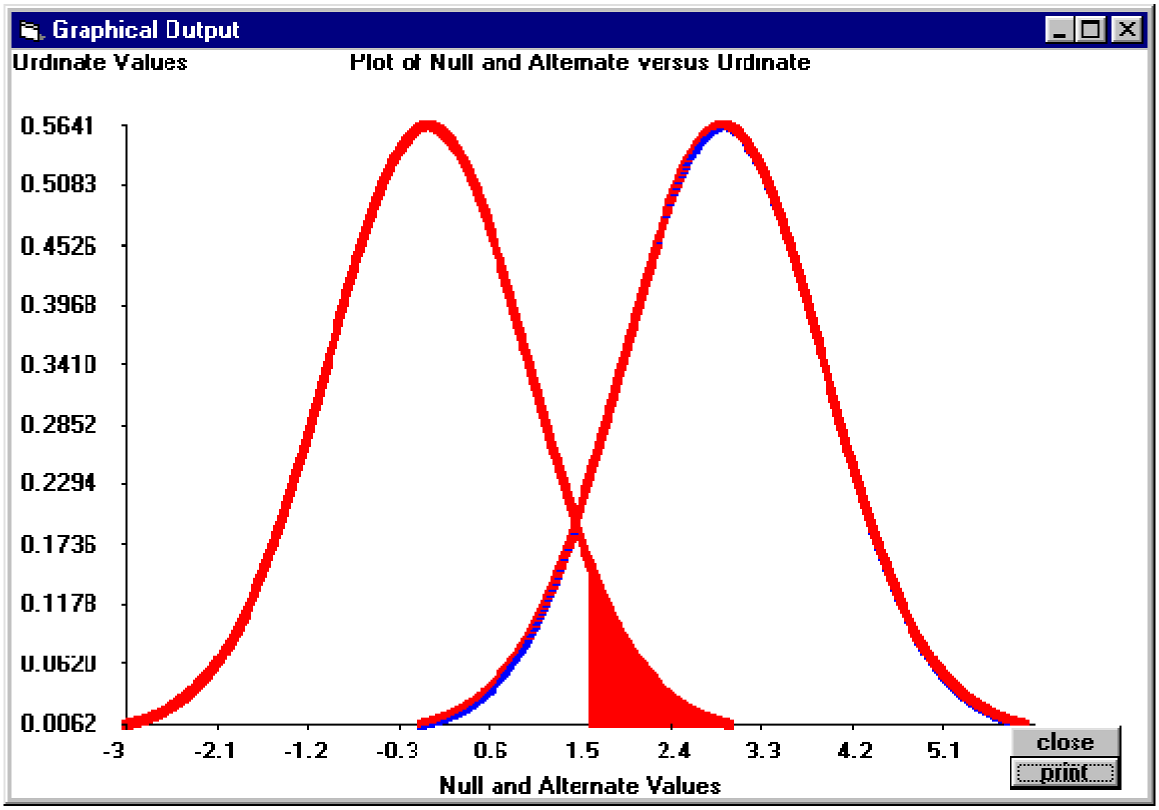


Fig. 3.4 Null and Alternate Hypotheses for Sample Means

This Fig. represents two population distributions of means for a variable. The distribution on the left represents the null hypothesized distribution. The distribution on the right represents an alternate hypothesis, that is, the hypothesis that a sample mean obtained is representative of a population in which the mean differs from the null distribution mean be a given difference D. The area of this latter distribution to the left of the shaded alpha area of the left curve and designated as ß represents the chance occurrence of a sample falling within the region of acceptance of the null hypothesis, even when drawn from the alternate hypothesized distribution. The score value corresponding to the critical mean value for this alternate distribution is:

$$\bar{X}_C = \sigma_{\bar{X}} z_{\beta} + \mu_1 \qquad (3.14)$$

Since formulas (1) and (2) presented above are both equal to the same critical value for the mean, they are equal to each other! Hence, we may solve for N, the sample size required in the following manner:

$$\sigma_{\bar{X}} z_{\alpha} + \mu_0 = \sigma_{\bar{X}} z_{\beta} + \mu_1 \qquad (3.15)$$

where $\mu_1 = \mu_0 - D$

and $\sigma_{\bar{X}} = \sigma_X / \sqrt{N}$ (3.16)

Therefore,

$$(\sigma_X / \sqrt{N})\, z_\alpha + \mu_0 = (\sigma_X / \sqrt{N})\, z_ß + \mu_1 \tag{3.17}$$

or $$\mu_1 - \mu_0 = (\sigma_x / \sqrt{N}) z_\alpha - (\sigma_x / \sqrt{N}) z_\beta) \tag{3.18}$$

or $$D = \sigma_x / \sqrt{N}\, (z_\alpha - z_\beta) \tag{3.19}$$

or $$\sqrt{N} = (\sigma_x / D)(z_\alpha - z_\beta) \tag{3.20}$$

Note: z_β is a negative value in the above drawing because we are showing an alternative hypothesis above the null hypothesis. For an alternative hypothesis below the null, the result would yield an equivalent formula.

By squaring both sides of the above equation, we have an expression for the sample size N required to maintain both the selected α rate and ß rate of errors, that is

$$N = \frac{\sigma_X^2}{D^2} (z_\alpha + z_ß)^2 \tag{3.21}$$

To demonstrate this formula (4) let us use the previous example of the teacher's experiment concerning a potentially superior teaching method. Assume that the teachers have agreed that it is important to contain both Type I error (α) and Type II error (ß) to the same value of .05. We may now determine the number of students that would be required to teach under the new teaching method and test. Remember that we wished to be sensitive to a difference between the population mean of 50 by at least 4 points in the positive direction only, that is, we must obtain a mean of at least 54 to have a meaningful difference in the teaching method. Since this is a "one-tailed" test, α will be in only one tail of the null distribution. The z score which corresponds to this α value is 1.645. Similarly the value of z corresponding to the ß level of .05 is also 1.645. The sample size is therefore obtained as

$$N = \frac{10^2}{4^2} (1.645 + 1.645)^2$$

$$= (100/16)\,(3.29)^2 = (100/16) * 10.81 = 67.65$$

or approximately 68 students.

Clearly, to provide control over both Type I and Type II error, our research is going to require a larger sample size than originally anticipated! In this situation, the teacher could simply repeat the teaching with his new method with additional sections of students or accept a higher Type II error.

It is indeed a sad reflection on much of the published research in the social sciences that little concern has been expressed for controlling Type II error. Yet, as we have seen, Type II error can often lead to more devastating costs or consequences than the Type I error which is usually specified! Perhaps most of the studies are restricted to small available (non-random) samples, or worse, the researcher has not seriously considered the costs of the types of error. Clearly, one <u>can</u> control both types of error and there is little excuse for not doing so!

Confidence Intervals for a Sample Mean

When a mean is determined from a sample of scores, there is no way to know anything certain about the value of the mean of the population from which the sample was drawn. We do know however sample means tend to be normally distributed about the population mean. If an infinite number of samples of size n were drawn at random, the means of those samples would themselves have a mean μ and a standard deviation of $\sigma / \sqrt{n}$. This standard deviation of the sample means is called the standard error of the mean because it reflects how much in error a sample mean is in estimating the population mean μ on the average. Knowing how far sample means tend to deviate from μ in the long run permits us to state with some confidence what the likelihood (probability) is that some interval around our single sample mean would actually include the population mean μ.

Since sample means do tend to be normally distributed about the population mean, we can use the unit-normal z distribution to make confidence statements about our sample mean. For example, using the normal distribution tables or programs, we can observe that 95 percent of normally distributed z scores have values between -1.96 and +1.96. Since sample means are assumed to be normally distributed, we may say that 95% of the sample means will surround the population mean μ in the interval of ± 1.96 the standard error of the means. In other words, if we draw a random sample of size n from a population of scores and calculate the sample mean, we can say with 95% confidence that the population mean is in the interval of our sample mean plus or minus 1.96 times the standard error of the means. Note however, that μ either is or is not in that interval. We cannot say for certain that μ is in the interval - only that we are some % confident that it is!

The calculation of the confidence interval for the mean is usually summarized in the following formula:

$$CI_{\%} = \bar{X} \pm z_{\%}\, \sigma_{\bar{X}} \qquad (3.22)$$

Using our previous example of this chapter, we can calculate the confidence interval for the sample mean of 55 and the standard error for the sample of 25 subjects = 2 as

$$CI_{95} = \bar{X} \pm (1.96)\, 2 \qquad (3.23)$$

$$= 51.08 \text{ to } 58.92$$

We state therefore that we are 95 percent confident that the actual population mean is between 58.92 and 51.08. Notice that the hypothesized mean (50) is not in this interval! This is consistent with our rejection of that null hypothesis. Had the mean of the null hypothesis been "captured" in our interval, we would have accepted the null hypothesis.

Another way of writing equation (5) above is

$$\text{probability } (\bar{X} - z_1\sigma_{\bar{X}} < \mu < \bar{X} + z_2\, \sigma_{\bar{X}}) = P \qquad (3.24)$$

where z_1 and z_2 are the z scores corresponding to the lower and upper values of the % confidence desired, and P is the probability corresponding to the % confidence.

For example we might have written our results of the teacher experiment as

probability [(55 - 1.96(2) < μ < 55 + 1.96(2)] = .95

or probability (51.08 < μ < 58.92) = .95

Frequency Distributions

A variable is some measure or observation of an attribute that varies from subject to subject. We are frequently interested in the shape of the distribution of the frequencies of objects who's scores fall in each category or interval of our variable. When the shape of the frequency distribution closely resembles that of a theoretical model of such distributions, we may utilize statistics developed for those theoretical distributions to describe our observations. We will examine some of the most common theoretical distributions. First, let us consider a simple Fig. representing the frequency of scores found in intervals of a classroom teacher's test. We will assume the teacher has administered a 20 item test to 80 students and has "plotted" the number of students obtaining the various total scores possible. The plot might look as follows:

```
Frequency
10                                  *
 9                              *      *
 8                          *             *
 7                       *                   *
 6
 5                    *                         *
 4
 3                 *                              *
 2               *                                   *
 1            *                                          *
 0      *  *                                                *  *
     ________________________________________________________________
      0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20
                          Total Test Score
```

Fig. 3.5 Sample Plot of Test Scores

We can also express the number of subjects in each score range as a proportion of the total number of observations. For example, we could divide each of the frequencies above by 80 (the number of observations) and obtain:

```
Proportion
.1250                               *
.1125                           *      *
.1000                       *             *
.0875                    *                   *
.0750
.0625                 *                         *
.0500
.0375              *                              *
.0250            *                                   *
.0125         *                                          *
.00  *  *  *                                                *  *
     ____________________________________________________________
   0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20
                           Total Test Score
```

Fig. 3.6 Sample Proportions of Test Scores

If the above distribution of the proportion of test scores at each possible value had been obtained on a very, very large number of cases in a population of subjects, we would refer to the proportions as probabilities. We would then be able to make statements such as "the probability of a student earning a score of 10 in the population is 0.01."

Sometimes we draw a Fig. that represents the cumulative frequencies divided by the total number of observations. For example, if we accumulate the frequencies represented in the previous Fig. the cumulative distribution would appear as:

```
Cum.Prob.
1.0                                                    *  *  *  *
0.9                                           *  *  *
0.8                                     *  *
0.7                                 *
0.6                              *
0.5
0.4                            *
0.3                         *
0.2                      *
0.1                *  *
0.0 *  *  *  *  *
    _____________________________________________________________
    0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
Total Test Score
```

Fig. 3.7 Sample Cumulative Probabilities of Test Scores

If the above 80 observations constituted the population of all possible observations on the 20 item test, we have no need of statistics to estimate population parameters. We would simply describe the mean and variance of the population values. If, on the other hand, the above 80 scores represents a random sample from a very, very large population of observations, we could anticipate that another sample of 80 cases might have a slightly different distribution appearance. The question may now be raised, what is a reasonable "model" for the distribution of the population of observations? There are clearly a multitude of distribution shapes for which the above sample of 80 scores might be reasonably thought to be a sample. Because we do not wish to examine all possible shapes that could be considered, we usually ask whether the sample distribution could be reasonably expected to have come from one of several "standard" distribution models. The one model having the widest application in statistics is called the "Normal Distribution". It is that model which we now examine.

The Normal Distribution Model

The Normal Distribution model is based on a mathematical function between the height of a probability curve for each possible value on the horizontal axis. Since the horizontal axis reflects measurement values, we must first translate our observations into "standard" units that may be used with any set of observations. The "z" score transformation is the one used, that is, we standardize our scores by dividing a scores deviation from the mean by the standard deviation of the scores. If we know the population mean and standard deviation, the transformation is

$$z_i = \frac{(X_i - \overline{X})}{\sigma_x} \tag{3.25}$$

If the population mean and standard deviation are unknown, then the sample estimates are used instead.

The Normal Distribution function (also sometimes called the Gaussian distribution function) is given by

$$h = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{3.26}$$

where h is the height of the curve at the value z and e is the constant 2.7182818.... .

To see the shape of the normal distribution for a large number of z scores, select the Analysis option and move the cursor to the Miscellaneous option. A second menu will appear. Click on the Normal Distribution Curve option. Values of h are drawn for values between approximately -3.0 to +3.0. It should be noted that the normal distribution actually includes values from -infinity to +infinity . The area under the normal curve totals 1.0. The area between any two z scores on the normal distribution therefore reflect the proportion (or probability) of z scores in that range. Since the z scores may be ANY value from -infinity to +infinity, the normal distribution reflects observations made on a scale considered to yield continuous scores.

The Median

While the mean is obtained as the average of scores in a distribution, it is not the only measure of "central tendency" or statistic descriptive of the "typical" score in a distribution. The *median* is also useful. It is the "middle score" or that value below which lies 50% of the remaining score values. The difference between the mean and median values is an indicator of how "skewed" are the distribution of scores. If the difference is positive (mean greater than the median) this would indicate that the mean is highly influenced by "extremely" high scores. If you plot the distribution of scores, there is typically a "tail" extending to the right (assuming the scores are arranged with low scores to the left and higher scores to the right.) We would say the distribution is positively skewed. When the distribution is negatively skewed the mean is less than the median. The median is highly useful for describing the typical score when the distribution is highly skewed. For example, the average income in the United States is much greater than the median income. A few millionaires (or billionaires) in the population skews the distribution. In this case, the median is more "representative" of the "typical" income.

Skew

The *skew* of a distribution is obtained as the third moment of the distribution. The first moment, the mean, is the average of the scores (sum of X's divided by the number of X's.) The second moment is the variance and is the average of the squared deviations from the mean. The third moment is the average of the cubed deviations from the mean. We can write this as:

$$\text{Skew} = \frac{\sum (X-\mu)^3}{N} \tag{3.27}$$

Kurtosis

A distribution may not only be skewed (not bell-shaped) but may also be "flatter" or more "peaked" than found in the normal curve. When a distribution is more flat we say that it is *platykurtik*. When it is more peaked we say it is *leptokurtik*. When it follows the typical normal curve it is described as *mesokurtik*. Kurtosis therefore describes the general height of the distribution across the score range. The kurtosis is obtained as the fourth moment about the mean. We can write it as:

$$\text{Kurtosis} = \frac{\sum (X-\mu)^4}{N} \tag{3.28}$$

The Binomial Distribution

Some observations yield a simple dichotomy that may be coded as 0 or 1. For example, you may draw a sample of subjects and observe the gender of each subject. A code of 1 may be used for males and 0 for females (or vice-versa). In a population of such scores, the proportion of observations coded 1 (P) is the mean (θ) of the scores. The population variance of dichotomous scores is simple θ(1-θ) or P(1-P). When a sample is drawn from a population of dichotomous scores, the sample mean, usually denoted simply as p, is an estimator of θ and the population variance is estimated by p(1-p). The probability of observing a specific number of subjects that would be coded 1 when sampling from a population in which the proportion of such subjects is P can be obtained from

$$X = \frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}$$

or simply

$$X = \binom{N}{n}p^n(1-p)^{N-n} \tag{3.29}$$

where X is the probability,
N is the size of the sample,
n is the number of subjects coded 1 and
P is the population proportion of objects coded 1.

The ! symbol in the above equation is the "factorial" operation, that is, n! means (1)(2)(3)....(n), the product of all integers up through n. Zero factorial is defined to be equal to 1, that is, 0! = 1.

For any sample of size N, we can calculate the probabilities of obtaining 0, 1, 2, ... , n values of the objects coded 1 when the population value is P. Once those values are obtained, we may also obtain the cumulative probability distribution. For example, assume you are sampling males and females from a population with a mean of 0.5, that is, the number of males (coded 1) equals the number of females (coded 0). Now assume you randomly select a sample of 10 subjects and count the number of males (n). The probabilities for n = 0, 1, ... , N are as follows:

No. Males Observed	Probability	Cumulative Probability
0	0.00097	0.00097
1	0.00977	0.01074
2	0.04395	0.05469
3	0.11719	0.17188
4	0.20508	0.37695
5	0.24609	0.62405
6	0.20508	0.82813
7	0.11719	0.94531
8	0.04395	0.98926
9	0.00977	0.99902
10	0.00097	1.00000

Now let us plot the above binomial distribution:

```
Probability
0.24-0.25                    *
0.22-0.23
0.20-0.21                *       *
0.18-0.19
0.16-0.17
0.14-0.15
0.12-0.13             *              *
0.10-0.11
0.08-0.09
0.06-0.07
0.04-0.05         *                       *
0.02-0.03
0.00-0.01 *   *                               *    *
          __________________________________________
          0   1   2   3   4   5   6   7   8   9   10
               Frequency of Males in Sample
            from a population with the number of
            males equal to the number of females
```

Fig. 3.8 Sample Probability Plot

The Poisson Distribution

The Poisson distribution describes the frequency with which discrete binomial events occur. For example, each child in a school system is either in attendance or not in attendance. The probability of each child being absent is, however, quite small. The probability of X children being absent from a school increases with the size of the school (n). Another example is in the area of school drop-outs. Each student may be considered to be either a drop-out or not a drop-out. The probability of being a drop-out student is usually quite small. The probability that X students out of n drop out over a given period of time may also be described by the Poisson distribution.

The Fig. below illustrates a representative Poisson distribution:

```
Frequency
 10 |
  9 |
  8 |
  7 |
  6 |           *
  5 |    *           *
  4 |*                   *
  3 |                       *
  2 |                          *
  1 |                            *
  0 |                                    *
    ----------------------------------------------
     0   1   2   3   4   5   6   7   8   9  10  11
          Course Dropouts Over 18 Week Period
                     (n = 50)
```

Fig. 3.9 A Poisson Distribution

The frequency (height) of the Poisson distribution is obtained from the following function:

$$f(x) = \frac{L^x e^{-L}}{x!} \tag{3.30}$$

where μ = L, the mean of the population distribution
and σ = L = the standard deviation of the population distribution

We note that when a variable (e.g. dropouts occurring) has a mean and standard deviation that are equal in the sample, the distribution may fit the Poisson model. In addition, it is important to remember that the variable (X) is a discrete variable, that is, only consists of integer values.

The Chi-Squared Distribution

In the field of statistics there is another important distribution that finds frequent use. The chi-squared statistic is most simply defined as the square of a normally distributed z score. Referring back to the paragraph on z scores, you will remember that is is obtained as

$$z_i = \frac{(X_i - \overline{X})}{\sigma_x} \tag{3.31}$$

that is, the deviation from the mean divided by the variance in the population of normally distributed scores. The z scores in an infinite population of scores ranges from -∞to +∞ . If we square randomly selected z scores, all resulting values are greater than or equal to zero. If we randomly select n z-scores, squaring each one, the sum of those squared z scores is defined as a Chi-Squared statistic with n degrees of freedom. Each time we draw a random sample of n z-scores and calculate the Chi-squared statistic, the value may vary from sample to sample. The distribution of these sample Chi-squared statistics follows the distribution density (height) function:

$$h = \frac{\chi^{\frac{n-2}{2}} e^{\frac{-\chi}{2}}}{\left[2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)\right]^{-1}} \tag{3.32}$$

where χ is the Chi-squared statistic,

n is the degrees of freedom,
e is the constant 2.7181... of the natural logarithm,
and Γ() is the gamma function.

In the calculation of the height of the chi-squared distribution, we encounter the gamma function (Γ). The gamma function is similar to another function, the factorial function (n!). The factorial of a number like 5, for example, is 5 x 4 x 3 x 2 x 1 which equals 80. The factorial however only applies to integer values. The gamma function however applies to continuous values as well as integer values. You can approximate the gamma function however by interpolating between integer values of the factorial. For example, the value of Γ(4) is equal to 3! or 3 x 2 x 1 = 6. In general, Γ(k-1) = k!

A sample distribution of Chi-squared statistics with 4 degrees of freedom is illustrated below

```
     .25

     .20
                  *    *
h    .15        *           *
              *                *
     .10     *                      *
            *                          *
     .05   *                                  *
           *                                        *
     .00   *                                                     *
         _________________________________________________________
           0    1    2    3    4    5    6    7    8    9   10   11
```

Fig. 3.10 Chi-squared Distribution with 4 Degrees of Freedom

The F Ratio Distribution

Another sample statistic which finds great use in the field of statistics is the F statistic. The F statistic may be defined in terms of the previously defined Chi-squared statistic. It is the ratio of two independent chi-squares, each of which has been divided by its degrees of freedom, that is

$$F_{(n_1,n_2)} = \frac{\dfrac{\chi^2}{n_1}}{\dfrac{\chi^2}{n_2}} \tag{3.33}$$

where χ^2 is the chi-squared statistic, and
n_1 and n_2 are the degrees of freedom for the numerator
and denominator chi-squares.

As before, we can develop the theoretical model for the sampling distribution of the F statistic. That is, we assume we repeatedly draw independent samples of n_1 and n_2 normally distributed z-scores, square each one, sum them up in each sample, and form a ratio of the two resulting chi-squared statistics. The height (density) function is given as

$$h = \frac{\Gamma\left[\dfrac{n_1+n_2}{2}\right] n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} \cdot \frac{F^{\frac{n_1-2}{2}}}{\left[n_1 F + n_2\right]^{\frac{n_1+n_2}{2}}} \tag{3.34}$$

where F is the sample statistic,
n_1 and n_2 are the degrees of freedom, and
Γ() is the gamma function described in the previous paragraph.

An example of the distribution of the F statistic for n_1 and n_2 degrees of freedom may be generated using the Distribution Plots and Critical Value procedure from the Simulation menu in your LazStats package.

The "Student" t Test

The z statistic used to test hypotheses concerning sample means assumes the use of the normal distribution. However we have seen that the unbiased estimate of the standard deviation of the sample means is "adjusted" for the sample size, that is S / √(N-1). If N is large, the distribution we can normally assume the distribution of the means is normal. When N is small, the "fit" to the normal distribution is less likely. William Gosset (who published under the name "Student") developed the mathematics for distributions that differ for the size of N but approach the normal (Gaussian) distribution as N increases in size. We obtain our statistic t in the same manner that we did for the z tests but instead of using the normal distribution, we use the t distribution. This distribution is described by the following equation:

$$y = \frac{C}{[1 + (t^2 / df)]^{(df+1)/2}} \tag{3.35}$$

Where

$$C = \frac{[(df - 1) / 2]!}{\sqrt{(n\,df)}\,[(df - 2) / 2]!}$$ Note: df is a single value for "degrees of freedom"

Shown below are two t distribution plots, the first with 2 degrees of freedom and the second with 100 degrees of freedom:

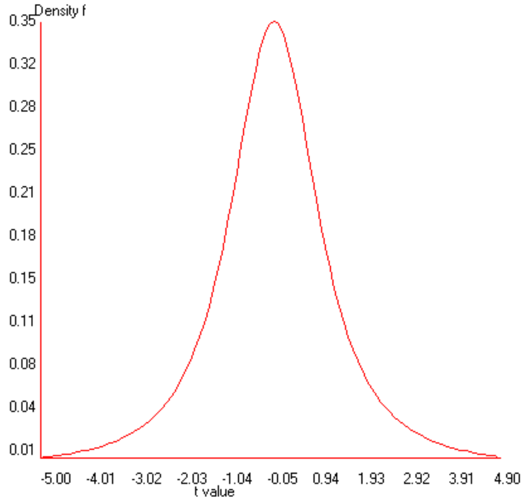


Fig. 3.11 t Distribution with 2 Degrees of Freedom

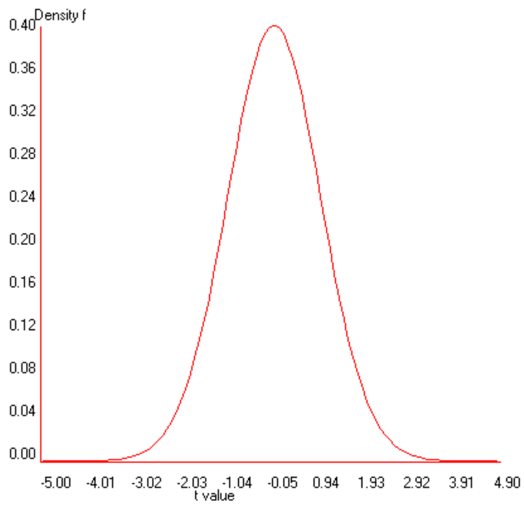


Fig. 3.12 t Distribution with 100 Degrees of Freedom

If you examine the density (height) of the curve on each of the above plots, you will see that the density is much greater for the plot with only 2 degrees of freedom. The "tails" of the t distribution are greater as the degrees of freedom decrease. If one is testing a hypothesis at the alpha level of say 0.05, it will take a larger value of t in a t-test in comparison to a z test to be significant for the smaller samples! The degrees of freedom for the t-test will vary depending on the nature of the hypothesis being tested.

Comparisons

One Sample Tests

LazStats provides the ability to perform tests of hypotheses based on a single sample. Typically the user is interested in testing the hypothesis that

1.a sample mean does not differ from a specified hypothesized mean,
2.a sample proportion does not differ from a specified population proportion,
3.a sample correlation does not differ from a specified population correlation, or

4.a sample variance does not differ from a specified population variance.

The One Sample Test for means, proportions, correlations and variances is started by selecting the Comparisons option under the Statistics menu and moving the mouse to the One Sample Tests option which you then click with the left mouse button. If you do this you will then see the specification form for your comparison as seen below. In this form there is a button corresponding to each of the above type of comparison. You click the one of your choice. There are also text boxes in which you enter the sample statistics for your test and select the confidence level desired for the test. We will illustrate each test. In the first one we will test the hypothesis that a sample mean of 105 does not differ from a hypothesized population mean of 100. The standard deviation is estimated to be 15 and our sample size is 20.

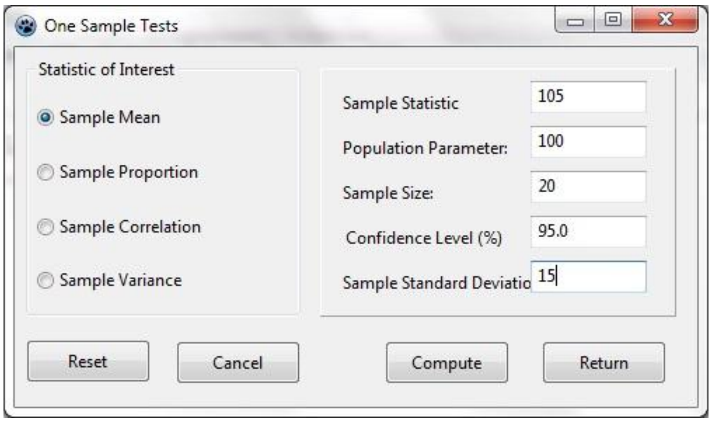


Fig. 3.13 Single Sample Tests Dialog For a Sample Mean

When we click the Continue button on the form we then obtain our results in an output form as shown below:

```
ANALYSIS OF A SAMPLE MEAN

Sample Mean = 105.000
Population Mean = 100.000
Sample Size = 20
Standard error of Mean =  3.354
t test statistic =  1.491 with probability  0.152
t value required for rejection =  2.093
Confidence Interval = (97.979,112.021)
```

We notice that our sample mean is "captured" in the 95 percent confidence interval and this would lead us to accept the null hypothesis that the sample is not different from that expected by chance alone from a population with mean 100.

Now let us perform a test of a sample proportion. Assume we have an elective high school course in Spanish I. We notice that the proportion of 30 students in the class that are female is only 0.4 (12 students) yet the population of high school students in composed of 50% male and 50% female. Is the proportion of females enrolled in the class representative of a random sample from the population? To test the hypothesis that the proportion of .4 does not differ from the population proportion of .5 we click the proportion button of the form and enter our sample data as shown below:

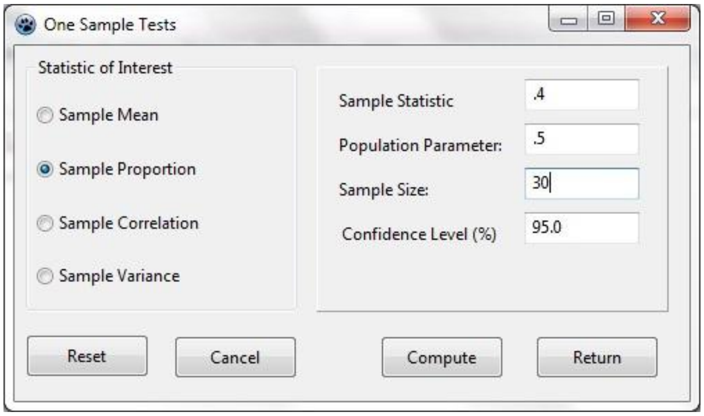


Fig. 3.14 One Sample Test for a Proportion

When we click the Continue button we see the results as shown below:

```
ANALYSIS OF A SAMPLE PROPORTION

Sample Proportion =  0.400
Population Proportion =  0.500
Sample Size = 30
Standard error of proportion =  0.091
z test statistic = -1.095 with probability > P =  0.863
z value required for rejection =  1.645
Confidence Interval = ( 0.221, 0.579)
```

We note that the z statistic obtained for our sample has a fairly high probability of occurring by chance when drawn from a population with a proportion of .5 so we are again led to accept the null hypothesis.

Now let us test a hypothesis concerning a sample correlation. Assume our Spanish teacher from the previous example has given two examinations to the 30 students enrolled in the course. The first is a standardized Spanish aptitude test and the second is a mid-term examination in the course. The teacher observes a correlation of 0.45 between the two examinations. In reading the literature which accompanies the standardized aptitude test the teacher notices that the validation study reported a correlation of 0.72 between the test and midterm examination scores in a very large sample of students. The teacher wonders if her observed correlation differs from that of the validation study. We enter our data in the Single Sample form as follows:

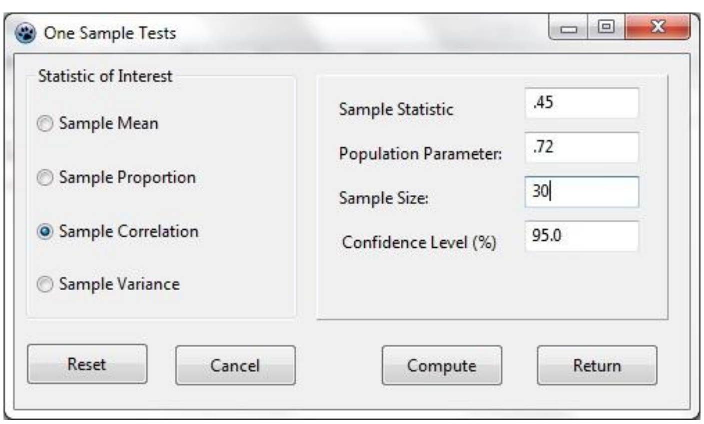


Fig. 3.15 One Sample Correlation Test

When the Continue button is pressed we obtain on the output form the following results:

```
ANALYSIS OF A SAMPLE CORRELATION

Sample Correlation =  0.450
Population Correlation =  0.720
Sample Size = 30
z Transform of sample correlation =  0.485
z Transform of population correlation =  0.908
Standard error of transform =  0.192
z test statistic = -2.198 with probability  0.014
z value required for rejection =  1.960
Confidence Interval for sample correlation = ( 0.107, 0.697)
```

Observing the small probability of the sample z statistic used to complete the test and noting that the population correlation is not in the 95% confidence interval for the sample statistic, our teacher reasonably rejects the null hypothesis of no difference and concludes that her correlation is significantly lower than that observed in the validation study reported in the test manual.

It occurs to our teacher in the above example that perhaps her Spanish students are from a more homogeneous population than that of the validation study reported in the standardized Spanish aptitude test. If that were the case, the correlation she observed might well be attenuated due to the differences in variances. In her class of thirty students she observed a sample variance of 25 while the validation study for the instrument reported a variance of 36. Let's examine the test for the hypothesis that her sample variance does not differ significantly from the "population" value. Again we invoke the One Sample Test from the Comparisons option of the Statistics menu and complete the form as shown below:

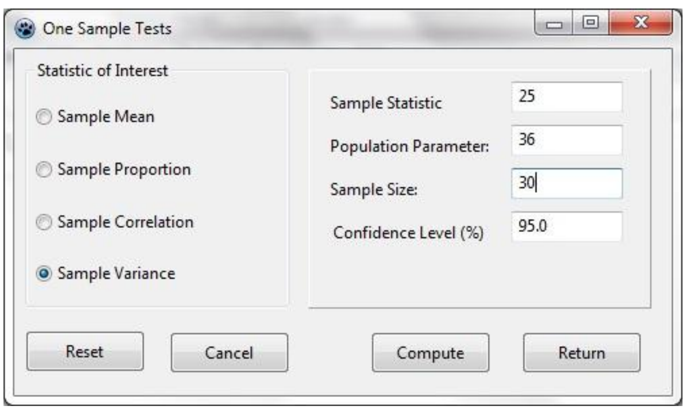


Fig. 3.16 One Sample Variance Test

Upon clicking the Continue button our teacher obtains the following results in the output form:

```
ANALYSIS OF A SAMPLE VARIANCE

Sample Variance = 25.000
Population Variance = 36.000
Sample Size = 30
Chi-square statistic = 20.139 with probability > chisquare =  0.889 and D.F.
= 29
Chi-square value required for rejection = 16.035
Chi-square Confidence Interval = (45.725,16.035)
Variance Confidence Interval = (15.856,45.215)
```

The chi-square statistic obtained leads our teacher to accept the hypothesis of no difference between her sample variance and the population variance. Note that the population variance is clearly within the 95% confidence interval for the sample variance.

Proportion Differences

A most common research question arises when an investigator has obtained two sample proportions. One asks whether or not the two sample proportions are really different considering that they are based on observations drawn randomly from a population. For example, a school nurse observes during the flu season that 13 eighth grade students are absent due to flu symptoms while only 8 of the ninth grade students are absent. The class sizes of the two grades are 110 and 121 respectively. The nurse decides to test the hypothesis that the two proportions (.118 and .066) do not differ significantly using the LazStats program. The first step is to start the Proportion Differences procedure by clicking on the Statistics menu, moving the mouse to the Comparisons option and the clicking on the Proportion Differences option. The specification form for the test then appears. We will enter the required values directly on the form and assume the samples are independent random samples from a population of eighth and ninth grade students.

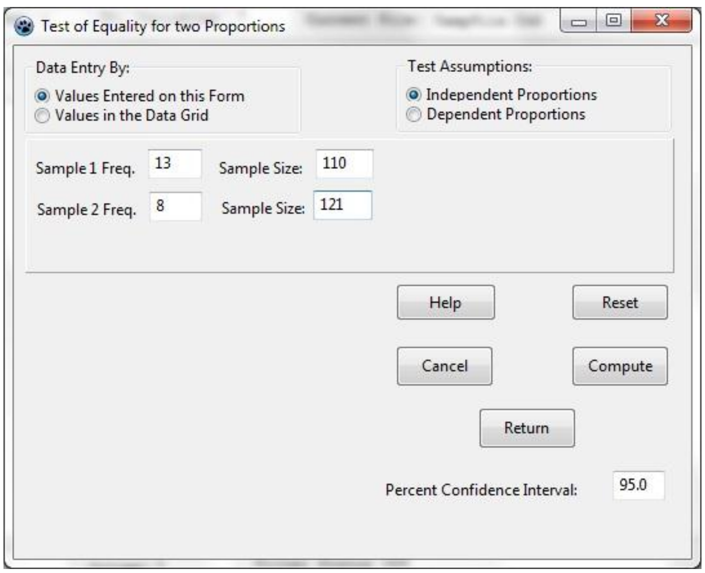


Fig. 3.17 Testing Equality of Two Proportions

When the nurse clicks the Continue button the following results are shown in the Output form:

```
COMPARISON OF TWO PROPORTIONS

Test for Difference Between Two Independent Proportions

Entered Values

Sample 1: Frequency =    13 for   110 cases.
Sample 2: Frequency =     8 for   121 cases.
Proportion 1 =     0.118, Proportion 2 =     0.066, Difference =     0.052
Standard Error of Difference =     0.038
Confidence Level selected = 95.0
z test statistic =     1.375 with probability = 0.0846
z value for confidence interval =  1.960
Confidence Interval: (   -0.022,    0.126)
```

The nurse notices that the value of zero is within the 95% confidence interval as therefore accepts the null hypothesis that the two proportions are not different than that expected due to random sampling variability.   What would the nurse conclude had the 80.0% confidence level been chosen?

If the nurse had created a data file with the above data entered into the grid such as:

CASE/VAR	FLU	GROUP
CASE 1	0	1
CASE 2	1	1
I.--		
CASE 110	0	1
CASE 111	0	2
–--		
CASE 231	1	2

then the option would have been to analyze data in a file.

In this case, the absence or presence of flu symptoms for the student are entered as zero (0) or one (1) and the grade is coded as 1 or 2. If the same students, say the eighth grade students, are observed at weeks 10 and 15 during the semester, than the test assumptions would be changed to Dependent Proportions. In that case the form changes again to accommodate two variables coded zero and one to reflect the observations for each student at weeks 10 and 15.

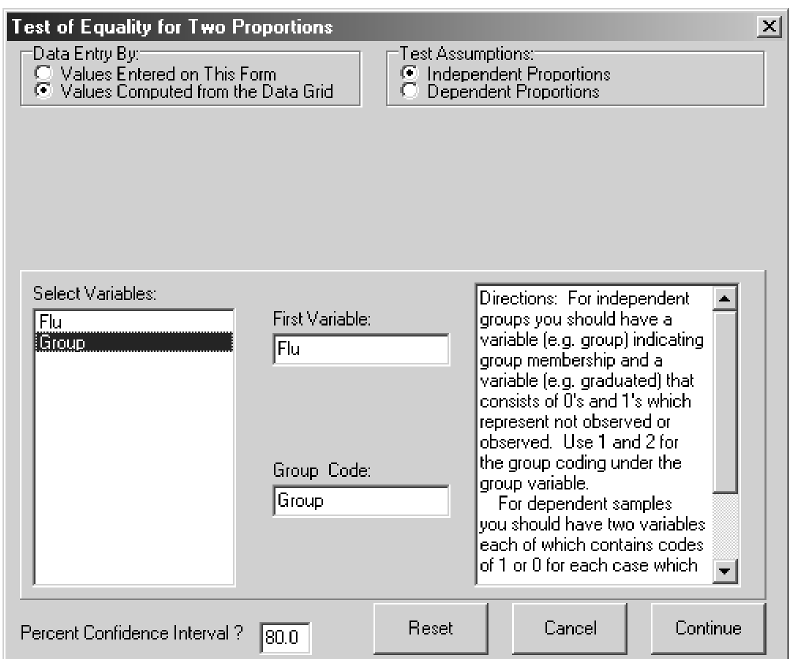


Fig. 3.18 Testing Equality of Two Independent Proportions (Grid Data)

Correlation Differences

When two or more samples are obtained, the investigator may be interested in testing the hypothesis that the two correlations do not differ beyond that expected due to random sampling variation. This test may be performed by selecting the correlation differences procedure in the comparison sub-menu of the statistics menu. The following form then appears:

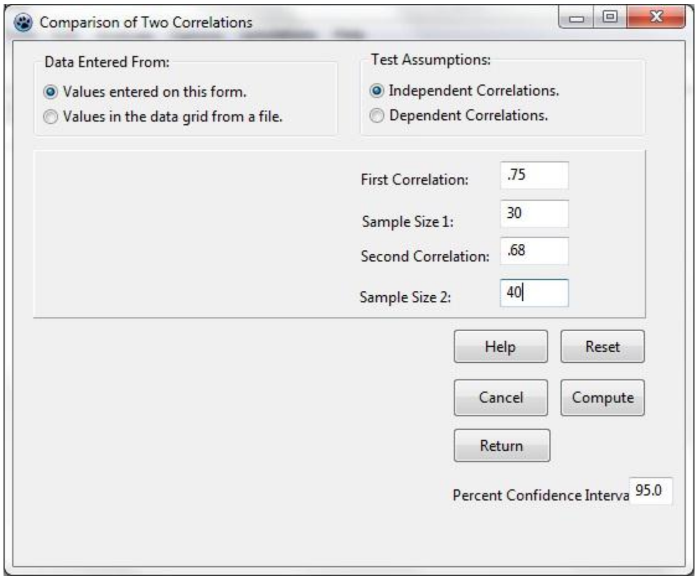


Fig. 3.19   Test of Difference Between Two Independent Correlations

Notice that the form above permit the user to enter the correlations directly on the form or to compute the correlations for two groups by reading the data from the data grid. In addition, the form permits the user to test the difference between correlations where the correlations are dependent. This may arise when the same two variables are correlated on the same sample of subjects at two different time periods or on samples which are "matched" on one or more related variables. As an example, let us test the difference between a correlation of .75 obtained from a sample with 30 subjects and a correlation of .68 obtained on a sample of 40 subjects. We enter our values in the "edit" fields of the form and click the continue button. The results appear below:

```
COMPARISON OF TWO CORRELATIONS

Correlation one =  0.750
Sample size one = 30
Correlation two =  0.680
Sample size two = 40
Difference between correlations =  0.070
Confidence level selected = 95.0
z for Correlation One =  0.973
z for Correlation Two =  0.829
z difference =  0.144
Standard error of difference =  0.253
z test statistic =  0.568
Probability > |z| =  0.285
z Required for significance =  1.960
Note: above is a two-tailed test.
Confidence Limits = (-0.338, 0.565)
```

The above test reflects the use of Fisher's log transformation of a correlation coefficient to an approximate z score. The correlations in each sample are converted to z's and a test of the difference between the z scores is performed. In this case, the difference obtained had a relatively large chance of occurrence when the null hypothesis is true (0.285) and the 95% confidence limit brackets the sample difference of 0.253. The Fisher z transformation of a correlation coefficient is

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

The test statistic for the difference between the two correlations is:

$$z_r = \frac{(z_{r_1} - z_{r_2}) - (z_{\rho_1} - z_{\rho_2})}{\sigma_{(z_{r_1} - z_{r_2})}}$$

where the denominator is the standard error of difference between two independent transformed correlations:

$$\sigma_{(z_{r_1} - z_{r_2})} = \sqrt{\left(\frac{1}{n_1 - 3}\right)\left(\frac{1}{n_2 - 3}\right)}$$

The confidence interval is constructed for the difference between the obtained z scores and the interval limits are then translated back to correlations. The confidence limit for the z scores is obtained as:

$$CI_{\%} = (z_{r_1} - z_{r_2}) + /- z_{\%}\sigma_{(z_{r_1} - z_{r_2})}$$

We can then translate the obtained upper and lower z values using:

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

For the test that two dependent correlations do not differ from zero we use the following t-test:

$$t = \frac{(r_{xy} - r_{xz})\sqrt{(n-3)(1 + r_{yz})}}{\sqrt{2(1 - r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + 2r_{xy}r_{xz}r_{yz})}}$$

Tests for Two Means

t-Tests

Among the comparison techniques the “Student” t-test is one of the most commonly employed. One may test hypotheses regarding the difference between population means for independent or dependent samples which meet or do not meet the assumptions of homogeneity of variance. To complete a t-test, select the t-test option from the Comparisons sub-menu of the Statistics menu. You will see the form below:

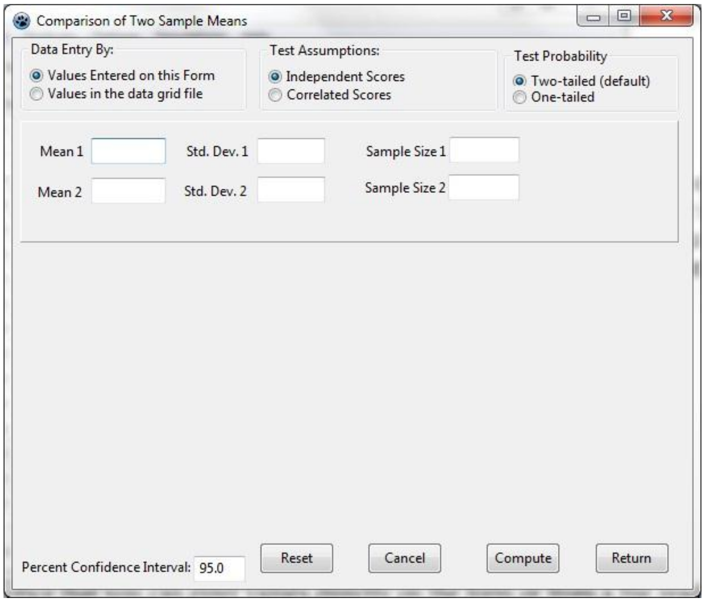


Fig. 3.20 Dialog Form For The Student t-Test

Notice that you can enter values directly on the form or from a file read into the data grid. If you elect to read data from the data grid by clicking the button corresponding to “Values Computed from the Data Grid” you will see that the form is modified as shown below.

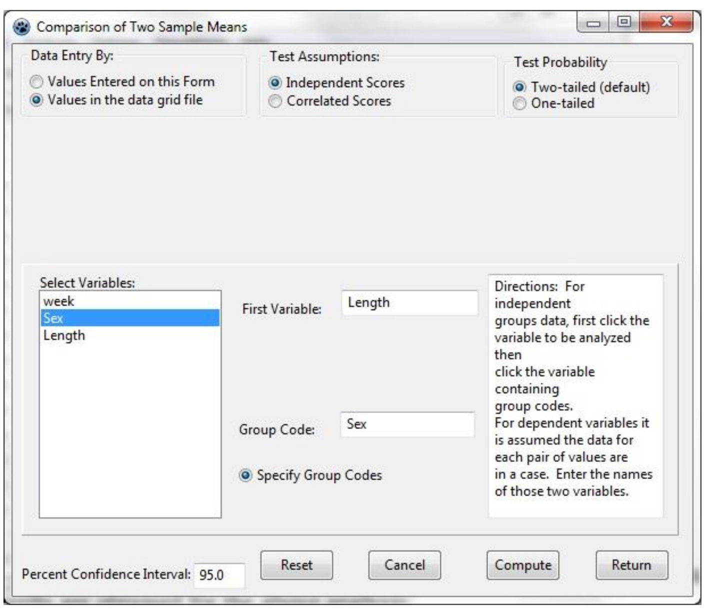


Fig. 3.21 Student t-Test For Data in the Data Grid

We will analyze data stored in the Hinkle411.LAZ file.

Once you have entered the variable name and the group code name you click the Continue button. The following results are obtained for the above analysis:

```
COMPARISON OF TWO MEANS

Variable  Mean       Variance  Std.Dev.  S.E.Mean  N
Group 1      31.00      67.74      8.23      1.68  24
Group 2      25.75      20.80      4.56      0.93  24
```

```
Assuming = variances, t =    2.733 with probability = 0.0089 and  46 degrees of freedom
Difference =     5.25 and Standard Error of difference =     1.92
Confidence interval = (    1.38,    9.12)
Assuming unequal variances, t =    2.733 with probability = 0.0097 and 35.91 degrees of freedom
Difference =     5.25 and Standard Error of difference =     1.92
Confidence interval = (    1.35,    9.15)
F test for equal variances =    3.256, Probability = 0.0032F test for equal variances =    3.256,
Probability = 0.0032

NOTE: t-tests are two-tailed tests.
```

The F test for equal variances indicates it is reasonable to assume the sampled populations have unequal variances hence we would report the results of the test assuming unequal variances. Since the probability of the obtained statistic is rather small (0.01), we would likely infer that the samples were drawn from two different populations. Note that the confidence interval for the observed difference is reported.

Chapter 4. The Product Moment Correlation

It seems most living creatures observe relationships, perhaps as a survival instinct. We observe signs that the weather is changing and prepare ourselves for the winter season. We observe that when seat belts are worn in cars that the number of fatalities in car accidents decrease. We observe that students that do well in one subject tend to perform will in other subjects. This chapter explores the linear relationship between observed phenomena.

If we make systematic observations of several phenomena using some scales of measurement to record our observations, we can sometimes see the relationship between them by "plotting" the measurements for each pair of measures of the observations. As a hypothetical example, assume you are a commercial artist and produce sketches for advertisement campaigns. The time given to produce each sketch varies widely depending on deadlines established by your employer. Each sketch you produce is ranked by five marketing executives and an average ranking produced (rank 1 = best, rank 5 = poorest.) You suspect there is a relationship between time given (in minutes) and the average quality ranking obtained. You decide to collect some data and observe the following:

Average Rank (Y)	Minutes (X)
3.8	10
2.6	35
4.0	5
1.8	42
3.0	30
2.6	32
2.8	31
3.2	26
3.6	11
2.8	33

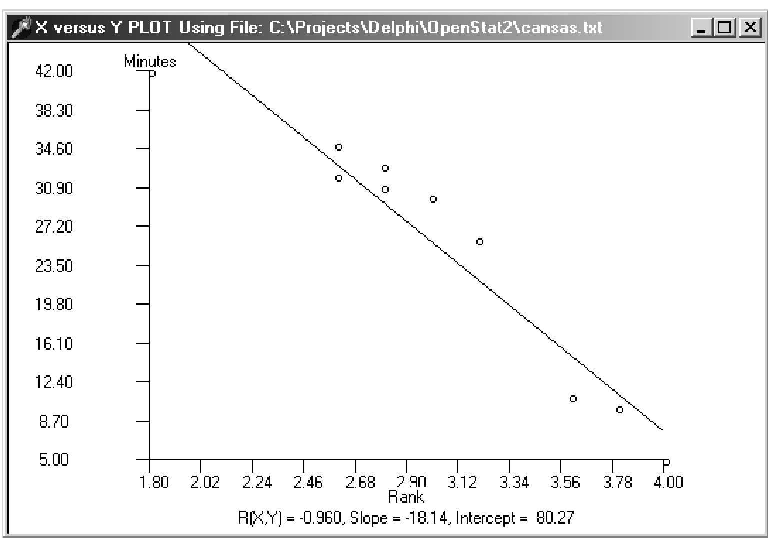


Fig. 4.1 A Negative Correlation Plot

Testing Hypotheses for Relationships Among Variables: Correlation

Scattergrams

While the mean and standard deviation of the previous chapter are useful for describing the central tendency and variability of the measures of a single variable, there are frequent situations in which it is desirable to obtain an index of how values measured on TWO variables tend to vary in the same or opposite directions. This "co-variability" of two variables may be visually represented by means of a Scattergram, for example, the Fig. below represents a scattergram of individual's scores on two variables, X and Y.

```
Scattergram of Two Variables

  14 |                                                     *
  13 |                                                 *     *
V 12 |                                             *   *  *   *
A 11 |                                        *  *    *     *
R 10 |                               *       *    *  *    *
I  9 |                                **  ***    *    *  *
A  8 |                             *     *    *    *  *
B  7 |                        *  *     *     *    *
L  6 |                   *    *    *       *   *
E  5 |              *     *     *      *    *
   4 |                 *     *      *     *
Y  3 |             *     *    *    *
   2 |              *  *    *
   1 |             *     *
_____|______|______|______|______|______|______|______|______|______|___
     0     1     2     3     4     5     6     7     8     9
VARIABLE X
```

Fig. 4.2 Scattergram of Two Variables

In the above Fig., each asterisk (*) represents a subject's position on two scales of measurement - on the X scale and on the Y scale. We can observe that subjects with larger X score values *tend* to have larger Y score values.

Now consider a set of score pairs representing measurements on two variables, College Grade Point Average (GPA) and Perceptions of Inadequacy (PI). The Fig. below the data represents the scattergram of subject scores.

Subject	GPA	PI
1	3.8	10
2	2.6	35
3	4.0	5
4	1.8	42
5	3.0	30
6	2.6	32
7	2.8	31
8	3.2	26
9	3.6	11
10	2.8	33

```
     SCATTERGRAM OF GPA VERSUS PI

GPA

4.0  *
3.8        *
3.6         *
3.4
3.2                           *
3.0                               *
2.8                                 * *
2.6                                  *  *
2.4
2.2
2.0
1.8                                              *
1.6
1.4__________________________________________________________
     5    10    15    20    25    30    35    40    45    50
                   Perceptions of Inadequacy
```

Fig. 4.3 Scattergram of a Negative Relationship

In this example there is a negative relationship between the two variables, that is, as a subject's perceptions of inadequacy increase, there is a corresponding decrease in grade point average! (The data are hypothetical if you hadn't guessed).

Many variables, of course, may not be related at all. In the following scattergram, there is no systematic co-variation between the two variables:

```
          Scattergram of Happiness and Wealth

Happiness

10                           *
 9                    *           *
 8                 *     *       *    *
 7                    *  *         *     *
 6           *    **    *     *       *  *
 5             *     *    *     *    * *   *
 4                *        *      *
 3                   *         *        *
 2                     *      *   *
 1                        *
 0_________________________________________________________
 0     1     2     3     4     5     6     7     8     9
          Wealth Measured as Thousands of Dollars in
                    a Checking Account
```

Fig. 4.4 Scattergram of Two Variables with Low Relationship

A simple way to construct an index of the relationship between two variables might be to simply average the product of the score pairs for the individuals. Unfortunately, the size of this index would vary as a function of the size of the numbers yielded by our measurement scales. We wouldn't be able to compare the index we obtained for, say, grade point averages in high school and college with the index we would obtain for college grade point averages and beginning salaries! On the other hand, an average of score products might be useful if we first transformed all of our measurements to a COMMON scale of measurement. In fact, this is just what Pearson did! By converting scores to a scale of measurement such that the mean score is always zero and the standard deviation of the scores on a variable is always 1.0, he was able to produce an index which, for any pair of variables, always varies between -1.0 and +1.0 !

Transformation to z Scores

We define a z score as a simple linear transformation of raw scores which involves the formula

$$z_i = \frac{\left(X_i - \overline{X}\right)}{s_x} \tag{4.1}$$

where z_i is the z score for an individual, X_i the individual's raw score and S_x is the standard deviation of the set of X scores.

When we have a pair of scores for each individual, we must adopt some method for differentiating between the two variables. Often we simply name the variables X and Y or X_1 and X_2. For the case of simple correlation discussed in this section, we will adopt the first method, i.e., the use of X and Y. We will use the second method when we start to deal with three or more variables at the same time.

The Pearson Product-Moment correlation is defined as

$$r_{x,y} = \frac{\sum_{i=1}^{N} z_{x_i} z_{y_i}}{N} \tag{4.2}$$

that is, the average of z score products for the N objects or subjects in our sample. Note, we have used the BIASED standard deviation in our z score transformations (divided by N, not N-1).

Now let us see how we apply the above formula in obtaining a coefficient of correlation for the above scattergram. First, we must transform our raw scores (Y and X) to z scores. To do this we must obtain the mean and standard deviation for each variable. In the Fig. below we have obtained the mean and standard deviation of each variable, obtained the deviation of each score from the respective mean, and finally, divided each deviation score by the corresponding standard deviation. We have also shown the product of the z scores for both the X and Y variables. It is this product of z scores which, when averaged, yields the *product-moment correlation* coefficient!

case	Y	X	$(Y_i - \bar{Y})$	$(X_i - \bar{X})$	$_y z_i$	$_x z_i$	$_y z_i \; _x z_i$
1	3.8	10	.78	- 15.5	1.253	-1.318	-1.651
2	2.6	35	- .42	9.5	- .675	.808	- .545
3	4.0	5	.98	- 20.5	1.574	-1.743	-2.743
4	1.8	42	-1.22	16.5	-1.960	1.403	-2.750
5	3.0	30	.02	4.5	.032	.383	.012
6	2.6	32	- .42	6.5	- .675	.553	- .373
7	2.8	31	- .22	5.5	- .353	.468	- .165
8	3.2	26	.18	0.5	.289	.043	.012
9	3.6	11	.58	- 14.5	.932	-1.233	-1.149
10	2.8	33	- .22	7.5	- .353	.638	- .225

$$\sum_{i=1}^{N} Y_i = 30.2 \qquad \bar{Y} = 3.02$$

$$\sum_{i=1}^{N} Y_i^2 = 95.08 \qquad S_y^2 = \sum_{i=1}^{N} Y^2 / N - \bar{Y}^2$$

$$= 9.508 - 9.1204 = 0.3876$$

$$\text{and } S_y = 0.62257529$$

$$\sum_{i=1}^{N} X_i = 255.0 \qquad \bar{X} = 25.5$$

$$\sum_{i=1}^{N} X_i^2 = 7885.0 \qquad S_x^2 = \sum_{i=1}^{N} X^2 / N - \bar{X}^2$$

$$= 788.5 - 650.25 = 138.25$$

$$\text{and } S_x = 11.757976$$

$$r_{yx} = \Sigma \; _y z_i \; _x z_i \; / \; N = -9.577 / 10 = -.958$$

The above method for obtaining the product-moment correlation is quite laborious and it is easy to make arithmetic mistakes and rounding errors. Let's look for another way which does not require actually computing the z scores for each variable. First, let us substitute the definition of the z scores in the formula for the correlation:

$$r_{x,y} = \frac{\sum_{i=1}^{N} z_{x_i} z_{y_i}}{N} = \frac{\sum_{i=1}^{N}\left(\frac{Y_i - \overline{Y}}{S_y}\right)\left(\frac{X_i - \overline{X}}{S_x}\right)}{N} = \frac{\sum_{i=1}^{N}\left[X_i Y_i - Y_i \overline{X} - \overline{Y} X_i + \overline{Y}\,\overline{X}\right]}{N S_y S_x} \tag{4.3}$$

or

$$r_{x,y} = \frac{\sum_{i=1}^{N} Y_i X_i - \sum_{i=1}^{N} Y_i \overline{X} - \sum_{i=1}^{N} \overline{Y} X_i + \sum_{i=1}^{N} \overline{Y}\,\overline{X}}{N S_y S_x} = \frac{\sum_{i=1}^{N} Y_i X_i - \overline{X}\sum_{i=1}^{N} Y_i - \overline{Y}\sum_{i=1}^{N} X_i + N\overline{Y}\,\overline{X}}{N S_y S_x} \tag{4.4}$$

or

$$r_{x,y} = \frac{\frac{\sum_{i=1}^{N} Y_i X_i}{N} - \overline{X}\frac{\sum_{i=1}^{N} Y_i}{N} - \overline{Y}\frac{\sum_{i=1}^{N} X_i}{N} + \overline{Y}\,\overline{X}}{S_y S_x} = \frac{\frac{\sum_{i=1}^{N} Y_i X_i}{N} - \overline{X}\,\overline{Y} - \overline{Y}\,\overline{X} + \overline{Y}\,\overline{X}}{S_x S_y} = \frac{\frac{\sum_{i=1}^{N} Y_i X_i}{N} - \overline{X}\,\overline{Y}}{S_x S_y} \tag{4.5}$$

The last formula does not require us to use z scores at all. We only need to use raw X and Y scores! Since we have already learned to compute S_x and S_y in terms of raw scores, we can do a little more algebra manipulation of the above formula and obtain

$$r_{x,y} = \frac{N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} Y_i\right)\left(\sum_{i=1}^{N} X_i\right)}{\sqrt{\left[N\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2\right]\left[N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2\right]}} \tag{4.6}$$

This formula is particularly advantages in that it utilizes the sums and sums of squared scores and the sum of cross-products of the X and Y scores. In addition, it contains fewer divisions which reduces round-off error! Using the previous example, we would obtain:

case	Y	X	Y^2	X^2	YX
1	3.8	10	14.44	100	38.0
2	2.6	35	6.76	1225	91.0
3	4.0	5	16.00	25	20.0
4	1.8	42	3.24	1764	75.6
5	3.0	30	9.00	900	90.0
6	2.6	32	6.76	1024	83.2
7	2.8	31	7.84	961	86.8
8	3.2	26	10.24	676	83.2
9	3.6	11	12.96	121	39.6
10	2.8	33	7.84	1089	92.4
	30.2	255	95.08	7885	699.8

$$r_{x,y} = \frac{(10)(699.8) - (30.2)(255)}{\sqrt{\left[10(95.08) - (30.2)^2\right]\left[10(7885) - (255)^2\right]}} = \frac{6998 - 7701}{\sqrt{[950.8 - 912.04][78850 - 65025]}}$$

or

$$r_{x,y} = \frac{-703}{\sqrt{(38.76)(13825)}} = \frac{-703}{\sqrt{535857}} = \frac{-703}{732.02254} = -0.960 \qquad (4.7)$$

(approximately)

Notice that the product-moment correlation obtained by this method differs by approximately .002 obtained in the average of z score products method. The first method had much more round-off error due to our calculations only being carried out to the nearest thousandths. Our results by this second method are clearly more accurate, even for only ten cases!

If you use the unbiased estimates of variances, other formulas may be written to obtain the product-moment correlation coefficient. Remember we divide the sum of squared deviations about the mean by N-1 for the unbiased estimate of population variance. In this case the average of z-score products is also divided by N-1 and by substituting the definition of a z score for both X and Y we obtain:

$$r_{x,y} = \frac{C_{x,y}}{s_x s_y} \qquad (4.8)$$

where

$$C_{x,y} = \frac{\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N}}{N-1}\ , \qquad (4.9)$$

the covariance of x and y

and the unbiased estimates of variance are:

$$s_x^2 = \frac{\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2 / N}{N-1} \qquad (4.10)$$

$$s_y^2 = \frac{\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2 / N}{N-1} \qquad (4.11)$$

with

$s_x = \sqrt{s^2_x}$ and $s_y = \sqrt{s^2_y}$

To further understand and learn to interpret the product-moment correlation, LazStats provides a means of simulating pairs of data, plotting those pairs, drawing the “best-fitting line” to the data points and showing the marginal distributions of the X and Y variables. Go to the Simulation menu and click on the Bivariate Scatter Plot. The Fig. below shows a simulation for a population correlation of -.95 with population means and variances as shown. A sample of 100 cases are generated. Actual sample means and standard deviations will vary (as sample statistics do!) from the population values specified.

```
POPULATION PARAMETERS FOR THE SIMULATION
Mean X :=   100.000, Std. Dev. X :=    15.000
Mean Y :=   100.000, Std. Dev. Y :=    15.000
Product-Moment Correlation :=    -0.900
Regression line slope :=    -0.900, constant :=   190.000
SAMPLE STATISTICS FOR 100 OBSERVATIONS FROM THE POPULATION
Mean X :=    99.988, Std. Dev. X :=    14.309
Mean Y :=   100.357, Std. Dev. Y :=    14.581
Product-Moment Correlation :=    -0.915
Regression line slope :=    -0.932, constant :=   193.577
```

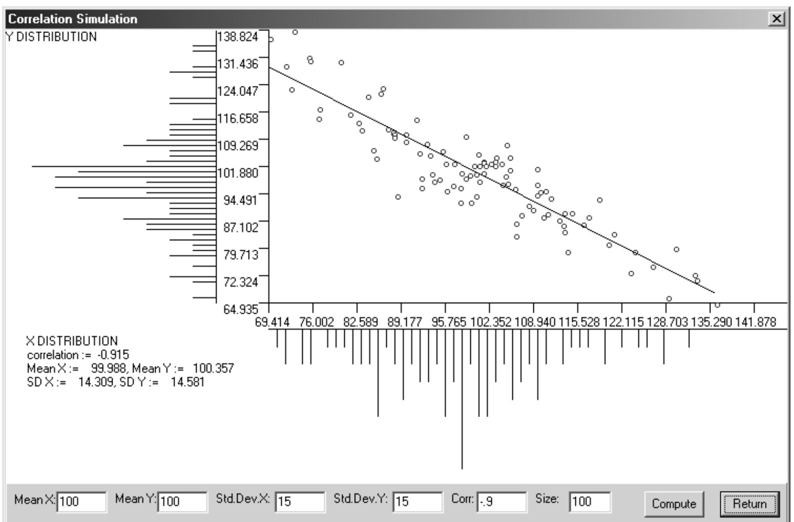


Fig. 4.5 A Simulated Negative Correlation Plot

Simple Linear Regression

The product-moment correlation discussed in the previous section is an index of the linear relationship between two continuous variables. But what is the nature of that linear relationship? That is, what is the slope of the line and where does the line intercept the vertical (Y variable) axis? This unit will examine the straight line "fit" to data points representing observations with two variables. We will also examine how this straight line may be used for prediction purposes as well as describing the relationship to the product-moment correlation coefficient.

To introduce the "straight line fit" we will first introduce the concept of "least-squares fit" of a line to a set of data points. To do this we will keep the number of X and Y score pairs small. Examine the Fig. below. It represents a set of 5 score pairs similar to those presented in the previous unit.

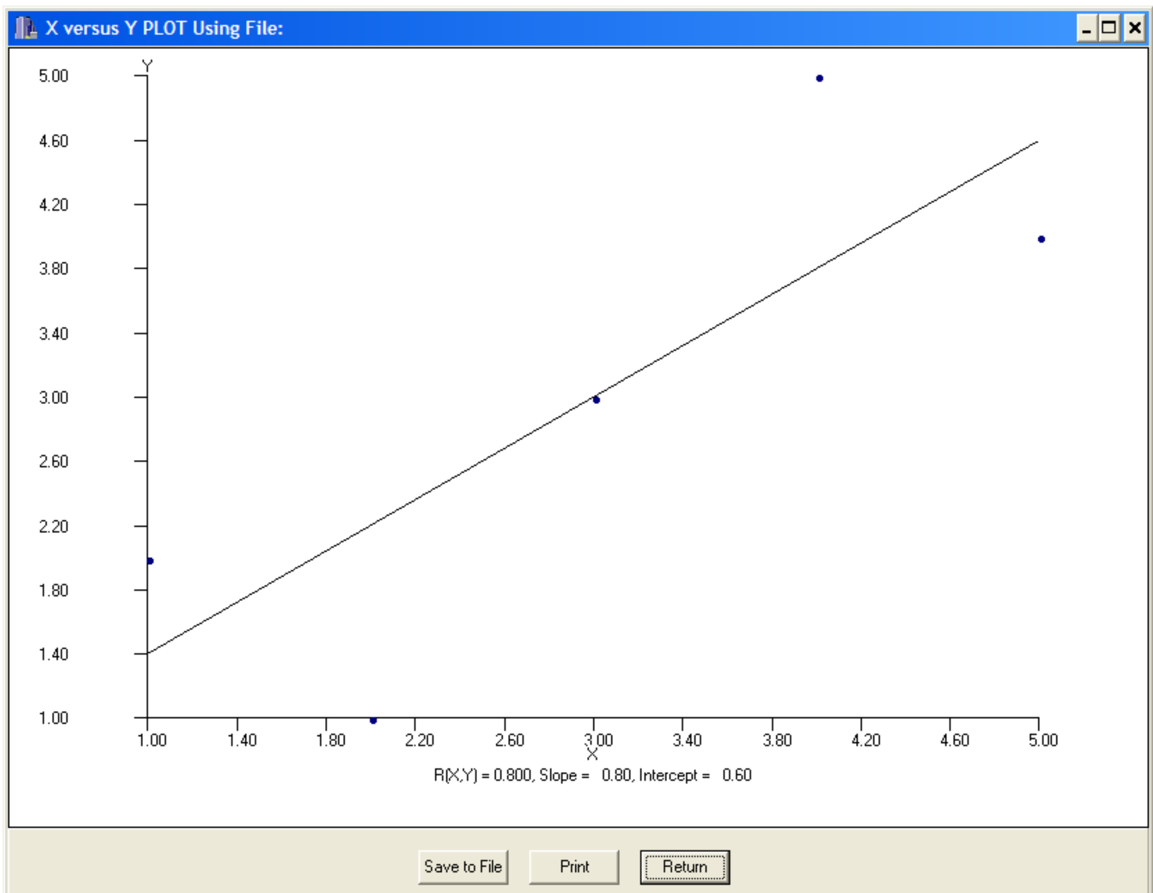


Fig. 4.6   X Versus Y Plot of Five Values

In the Fig., each point represents the intersection of X and Y score values for an observed case. Also shown is a line that represents the "best fitting line" to the data points:

```
Case      1   2   3   4   5
     -----------------
     X | 1   2   3   4   5
     Y | 2   1   3   5   4
     -----------------
```

The Least-Squares Fit Criterion

In regression analysis, we want to develop a formula for a straight line which optimally predicts each Y score from a given X score. For example, if Y is a student's College Grade Point Average (GPA) and X is the high school grade point average (HSGPA), we wish to develop an equation which will predict the GPA given the HSGPA. Straight line formulas generally are of the form

$$Y = BX + C \tag{4.12}$$

where B is the slope of the line,
and C is a constant representing the point where the line crosses the Y axis. This is also called the intercept.

In the Fig. below, B is the slope of the line (the number of Y units (rise) over 1 unit of X (run). C is the intercept where the line crosses the Y axis.

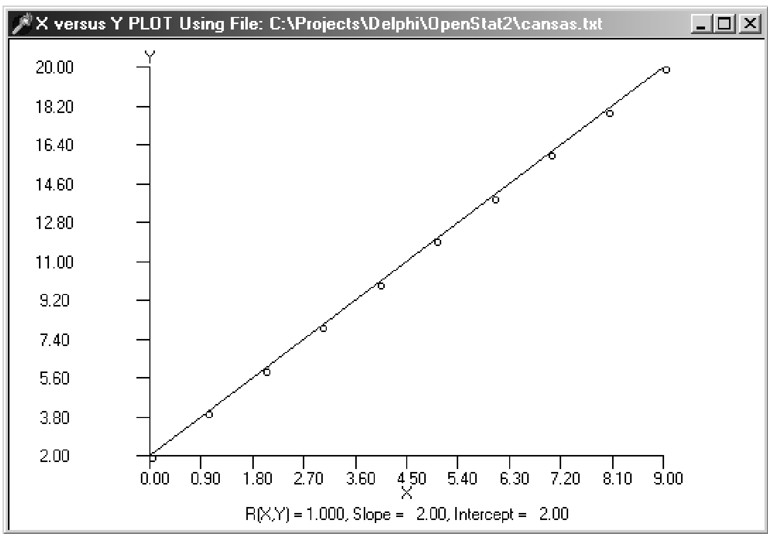


Fig. 4.7 Plot for a Correlation of 1.0

If X and Y scores are transformed to z scores using the transformations

$$z_x = (X_i - \bar{X}) / \sigma_x \tag{4.13}$$

and

$$z_y = (Y_i - \bar{Y}) / \sigma_y \tag{4.14}$$

then we may write for our prediction of the corresponding z_y scores

$$z_y' = bz_x + 0 \tag{4.15}$$

since the intercept is zero for z scores.

The Least-Squares criterion implies that the squared difference between each predicted score and actual observed score Y is a minimum. That is

$$\sum_{i=1}^{N} (z_y - z_y')^2 = \text{Minimum} \tag{4.16}$$

where z_y' is the predicted z_y score for an individual.

The problem is to obtain values of b such that the above statement is true. If we substitute bz_x for each z_y' in the above equation and expand, we get

$$\text{Min} = \sum_{i=1}^{N} [z_y - bz_x]^2 \tag{4.17}$$

$$= \sum_{i=1}^{N} (z_y^2 + b^2 z_x^2 - 2bz_y z_x)$$

$$= \sum_{i=1}^{N} z_y^2 + b^2 \sum_{i=1}^{N} z_x^2 - 2b \sum_{i=1}^{N} z_y z_x \tag{4.18}$$

In the mathematics called Calculus, it is learned that the first derivative of a function is either a minimum or a maximum. By taking the partial derivative of the above function Min (we will call it M) with respect to b, we get an equation which can be solved for b. This equation is set equal to zero and solved for b. The derivative of M with respect to be is:

$$\frac{\delta M}{\delta b} = 2b \sum z_x^2 - 2 \sum z_y z_x \tag{4.19}$$

Setting the derivative to zero and solving for b gives

$$0 = b \sum z_x^2 - \sum z_y z_x \tag{4.20}$$

or $$b = \sum z_y z_x / \sum z_x^2 \tag{4.21}$$

Since the sum of squared z scores is equal to N (if we use the biased standard deviation), we see that

$$b = \sum z_x z_y / N .$$

The product-moment correlation was earlier defined to be the average of z score products. Therefore, the slope of a regression line in z score form is simply

$$b = r_{xy}$$

The prediction equation is therefore

$$z_y' = r_{xy} z_x \tag{4.22}$$

To determine the values of B and C in the equation for raw scores, simply substitute the definition of z scores in the above equation, that is

$$\frac{(Y' - \bar{Y})}{s_y} = r_{xy} \frac{(X - \bar{X})}{s_x} \tag{4.23}$$

$$\text{or} \quad (Y' - \bar{Y}) = r_{xy} \frac{s_y}{s_x} (X - \bar{X}) \tag{4.24}$$

$$\text{or} \quad Y' = r_{xy} \frac{s_y}{s_x} X - r_{xy} \frac{s_y}{s_x} \bar{X} + \bar{Y} \tag{4.25}$$

Letting $B = r_{xy}(s_y / s_x)$, the last equation may be written

$$Y' = B X - (B \bar{X} - \bar{Y}) \tag{4.26}$$

To express the equation is the typical "straight line" equation, let

$$C = \bar{Y} - B \bar{X} \tag{4.27}$$

so that $$Y' = B X + C \tag{4.28}$$

To summarize, the least-squares criterion is met when the predicted scores for z_y or Y are obtained from

$$z_y' = r z_x \tag{4.29}$$

or $Y' = B X + C$ where $B = r_{xy} (s_y / s_x)$ and

$$C = \bar{Y} - B \bar{X} \tag{4.30}$$

The Variance of Predicted Scores

We can develop an expression for the variance of predicted scores z_y' or Y'. Using the definition of variance, we have

$$s^2_{Y'} = \frac{(Y' - \bar{Y})^2}{N} \tag{4.31}$$

By substituting the definition of Y', that is, BX + C, in the above equation, we could show that the variance of predicted scores is

$$s^2_{Y'} = r_{xy}^2 \, s_y^2 \tag{4.32}$$

That is, the variance of the predicted scores is the square of the product-moment correlation between X and Y times the variance of the Y scores. It is also useful to rewrite the above equation as

$$r_{xy}^2 = s^2_{Y'} / s_y^2 \ . \tag{4.33}$$

The square of the correlation is that proportion of total score variance that is predicted by X !

The Variance of Errors of Prediction

Just as we developed an expression for the variance of predicted scores above, we can also develop an expression for the variance of errors of prediction, that is, the variance of

$e_i = (Y_i - Y_i')$ for each score.

Again using the definition of variance we can write

$$s^2_{Y.X} = \frac{\Sigma (Y_i - Y_i')^2}{N} \tag{4.34}$$

This formula is biased due to estimating both the mean of X as well as the mean of Y in the population. For that reason the unbiased estimate is

$$s^2_{Y.X} = \frac{\Sigma e_i^2}{N - 2} \tag{4.35}$$

The square root of this variance is called the **standard error of estimate**. When we can assume the errors of prediction are normally distributed, it allows us to estimate a confidence interval for a given predicted score.

Rather than having to compute an error for each individual, the above formula may be translated into a more convenient computational form:

$$s^2_{Y.X} = s_y^2 (1 - r^2_{xy}) \frac{N - 1}{N - 2} \tag{4.36}$$

As an example in using the standard error of estimate, assume we have obtained a correlation of 0.8 between scores of X and Y for 40 subjects. If the variance of the Y scores is 100, then the variance of estimate is

$$s^2_{Y.X} = 100 \ (1.0 - 0.64) \ (19 / 18)$$

$$= 38$$

and

$$S_{Y.X} = \sqrt{38} = 6.1644$$

Using plus or minus 1 under the normal distribution, we can state that a predicted Y score would be expected to be in the interval ($Y' \pm 6.2$) approximately 68 percent of the time.

Testing Hypotheses Concerning the Pearson Product-Moment Correlation.

Hypotheses About Correlations in One Population

The product-moment correlation is an index of the linear relationship between two variables that varies between -1.0 and +1.0 with a value of 0.0 indicating no relationship. When obtaining pairs of X and Y scores on a sample of subjects drawn from a population, one can hypothesize that the correlation in the population does not differ from zero (0), i.e. H_0: $\Upsilon = 0$. The test statistic is:

$$t = \frac{r-\gamma}{S_r} \text{ with n-2 degrees of freedom, and} \tag{4.37}$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}} \tag{4.38}$$

As an example, assume a sample correlation r = 0.3 is obtained from a random selection of 38 subjects from a population of subjects. To test the hypothesis that the population correlation does not differ significantly from zero in either direction, we would obtain

$$S_r = \sqrt{\frac{1-.09}{38-2}} = 0.158989866 \tag{4.39}$$

and

$$t = r / S_r = .3 / 0.158989866 = 1.886912706 \tag{4.40}$$

With n-2 = 36 degrees of freedom, the t value obtained would be considered significant at the 0.05 level for a one-tailed test (r > 0), hence we would fail to retain the null hypothesis (reject).

Test That the Correlation Equals a Specific Value

The sampling distribution of the product-moment correlation is approximately normal or t distributed when sampled from a population in which the true correlation is zero. Occasionally, however, one wishes to test the hypothesis that the population correlation does not differ from some specified value ρ not equal to zero. The distribution of sample correlations from a population in which the correlation differs from zero is skewed, with the degree of skewness increasing as the population correlation differs from zero. It is possible to transform the correlations to a statistic which has a sampling distribution that is approximately normal in shape. The transformation, credited to Fisher, is:

$$z_r = 0.5 \log_e\left(\frac{1+r}{1-r}\right) \tag{4.41}$$

This statistic has a standard error of:

$$S_r = \sigma[1/(n-3)] \tag{4.42}$$

Using the above, a t-test for the hypothesis H_0:ρ=a can be obtained as

$$z = \frac{z_r - z_\rho}{S_r} \tag{4.43}$$

For example, assume we have obtained a sample correlation of r = 0.6 on 50 subjects and we wish to test the hypothesis that the population correlation does not differ from 0.5 in the positive direction. We would first transform both the sample and population correlations to the Fisher's z score and obtain:

$$z_r = .5\log_e[(1+.6)/(1-.6)] = 0.6931472 \tag{4.44}$$

and

$$z_\rho = .5\log_e[(1+.5)/(1-.5)] = 0.5493061 \tag{4.45}$$

Next, we obtain the standard error as

$$S_{z_r} = \sigma[1/(n-3)] = \sigma[1/(50-3)] = 0.145865 \tag{4.46}$$

Our test statistic is then

$$z = \frac{z_r - z_\rho}{S_{z_r}} = \frac{0.143841}{0.145865} = 0.986 \tag{4.47}$$

Approximately .16 of the area of the normal curve lies beyond a z of .986. We would retain our null hypothesis if our decision rule was for a probability of 0.05 or less in order to reject.

As for all of the sample statistics discussed so far, a confidence interval may be constructed. In the case of the Fisher's z transformation of the correlation, we first construct our interval using the z-transformed scores and then obtain the anti-log to express the interval in terms of product-moment correlations. For example, the 90% Confidence Interval for the above data is obtained as:

$$CI_{90} = z_r \pm 1.645(S_{z_r}) \tag{4.48}$$

$$= .693 \pm 1.645(.146) = .693 \pm 0.24$$

$$= (.453, .933)$$

and transforming the z_r intervals to r intervals gives

$$CI_{90} = (0.424, 0.732) \tag{4.49}$$

We converted the z_r values back to correlations using

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1} \tag{4.50}$$

Notice that the sample value of 0.6 is "captured" in the 90% Confidence Interval, thus verifying our one-tailed 0.05 test.

LazStats contains a procedure for completing a z test for data like that presented above.

Under the Statistics menu, move your mouse down to the Comparisons sub-menu, and then to the option entitled "One Sample Tests". When the form below displays, click on the Correlation button and enter the sample value .5, the population value .6, and the sample size 50. Change the confidence level to 90.0%.

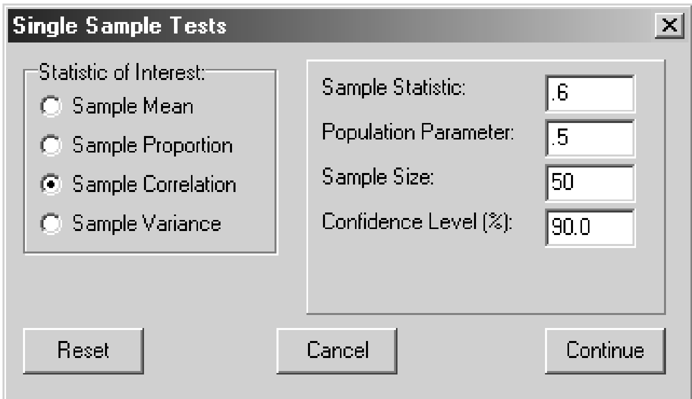


Fig. 4.8 Single Sample Tests Dialog Form

Shown below is the z-test for the above data:

```
ANALYSIS OF A SAMPLE CORRELATION

Sample Correlation =  0.600
Population Correlation =  0.500
Sample Size = 50
z Transform of sample correlation =  0.693
z Transform of population correlation =  0.549
Standard error of transform =  0.146
z test statistic =  0.986 with probability  0.838
z value required for rejection =  1.645
Confidence Interval for sample correlation = ( 0.425, 0.732)
```

Testing Equality of Correlations in Two Populations

When two populations have been sampled, a correlation between X and Y scores of each sample are often obtained. We may test the hypothesis that the product-moment correlation in the two populations are equal. If we assume the samples are independent, our test statistic may be obtained as

$$z = \frac{\left(z_{r_1} - z_{r_2}\right) - \left(z_{\gamma_1} - z_{\gamma_2}\right)}{S_{\left(z_{r_1} - z_{r_2}\right)}} \tag{3.59}$$

where

$$S_{\left(z_{r_1} - z_{r_2}\right)} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{3.60}$$

As an example, assume we have collected ACT Composite scores (a college aptitude test) and College Freshman Grade Point Average (GPA) scores for both men and women at a state university. We might hypothesize that in the population of men and women at this university, there is no difference between the correlation of GPA and ACT. Now pretend that a sample of 30 men yielded a correlation of .5 and that a sample of 40 women yielded a correlation of .6. Our test would yield:

z_r = 0.5493061 for the men,

z_r = 0.6931472 for the women, and

$$S_{(z_{r_1} - z_{r_2})} = \sqrt{\frac{1}{27} + \frac{1}{37}} = 0.253108798 \tag{3.61}$$

and the test value of

z = (0.5493061 - 0.6931472) / 0.253108798

= -0.568

which would not be significant.

The above test reflects the use of Fisher's log transformation of a correlation coefficient to an approximate z score. The correlations in each sample are converted to z's and a test of the difference between the z scores is performed. In this case, the difference obtained had a relatively large chance of occurrence when the null hypothesis is true (0.285) and the 95% confidence limit brackets the sample difference of 0.253. The Fisher z transformation of a correlation coefficient is

$$z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right) \tag{3.62}$$

The test statistic for the difference between the two correlations is:

$$z_r = \frac{(z_{r_1} - z_{r_2}) - (z_{\rho_1} - z_{\rho_2})}{\sigma_{(z_{r_1} - z_{r_2})}} \tag{3.63}$$

where the denominator is the standard error of difference between two independent transformed correlations:

$$\sigma_{(z_{r_1} - z_{r_2})} = \sqrt{\left(\frac{1}{n_1 - 3} \right) \left(\frac{1}{n_2 - 3} \right)} \tag{3.64}$$

The confidence interval is constructed for the difference between the obtained z scores and the interval limits are then translated back to correlations. The confidence limit for the z scores is obtained as:

$$CI_{\%} = (z_{r_1} - z_{r_2}) + / - z_{\%} \sigma_{(z_{r_1} - z_{r_2})} \tag{3.65}$$

We can then translate the obtained upper and lower z values using:

$$r = \frac{e^{2 z_r} - 1}{e^{2 z_r} + 1} \tag{3.66}$$

For the test that two dependent correlations do not differ from zero we use the following t-test:

$$t = \frac{\left(r_{xy} - r_{xz}\right)\sqrt{(n-3)\left(1+r_{yz}\right)}}{\sqrt{2\left(1 - r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + 2r_{xy}r_{xz}r_{yz}\right)}} \tag{3.67}$$

We would therefore conclude that, in the populations sampled, there is not a significant difference between the correlations for men and women. Using LazStats to accomplish the above calculations is rather easy. Under the Statistics menu move to the Comparisons sub-menu and further in that menu to the Two-Sample Tests sub-sub-menu. Click on the Independent Correlations option. Shown below are the results for the above data:

```
COMPARISON OF TWO CORRELATIONS

Correlation one =  0.500
Sample size one = 30
Correlation two =  0.600
Sample size two = 40
Difference between correlations = -0.100
Confidence level selected = 95.0
z for Correlation One =  0.549
z for Correlation Two =  0.693
z difference = -0.144
Standard error of difference =  0.253
z test statistic = -0.568
Probability > |z| =  0.715
z Required for significance =  1.960
Note: above is a two-tailed test.
Confidence Limits = (-0.565, 0.338)
```

Differences Between Correlations in Dependent Samples

Assume that three variables are available for a population of subjects. For example, you may have ACT scores, Freshman GPA (FGPA) scores and High School GPA (HSGPA) scores. It may be of interest to know whether the correlation of ACT scores with High School GPA is equal to the correlation of ACT scores with the Freshman GPA obtained in College. Since the correlations would be obtained across the same subjects, we have dependency between the correlations. In other words, to test the hypothesis that the two correlations r_{xy} and r_{xz} are equal, we must take into consideration the correlation r_{yz} . A t-test with degrees of freedom equal to N-3 may be obtained to test the hypothesis that $\Upsilon_{xy} = \Upsilon_{xz}$ in the population. Our t-test is constructed as

$$t = \frac{r_{x,y} - r_{x,z}}{\sqrt{\dfrac{2\left(1 - r_{x,y}^2 - r_{x,z}^2 - r_{y,z}^2 + 2r_{x,y}r_{x,z}r_{y,z}\right)}{(N-3)\left(1+r_{y,z}\right)}}} \tag{3.68}$$

Assume we have drawn a sample of 50 college freshman and observed:

r_{xy} = .4 for the correlation of ACT and FGPA, and

r_{xz} = .6 for the correlation of ACT and HSGPA, and

r_{yz} = .7 for the correlation of FGPA and HSGPA.

Then for the hypothesis that $\Upsilon_{xy} = \Upsilon_{xz}$ in the population of students sampled, we have

$$t = \frac{.4-.6}{\sqrt{\dfrac{2\left[1-.4^2-.6^2-.7^2+2(.4)(.6)(.7)\right]}{(50-3)(1+.7)}}} = \frac{-.2}{\sqrt{\dfrac{.652}{79.9}}} = \frac{-.2}{0.0903338} = 2.214 \tag{3.69}$$

This sample t value has a two-tailed probability of less than 0.05. If the 0.05 level were used for our decision process, we would reject the hypothesis of equal correlations of ACT with the high school GPA and the freshman college GPA. It would appear that the correlation of the ACT with high school GPA is greater than with College GPA in the population studied.

Again, LazStats provides the computations for the difference between dependent correlations as shown in the Fig. below:

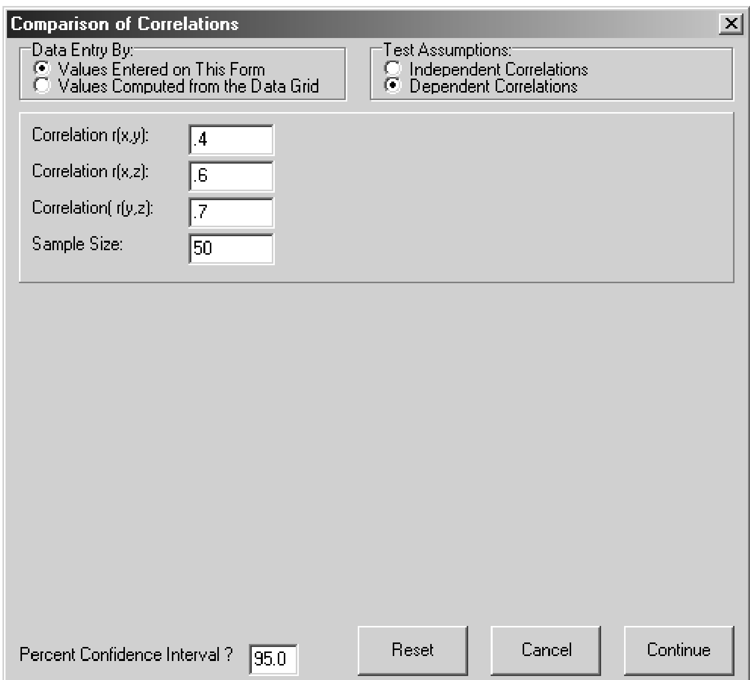

Comparison of Correlations
Data Entry By:
Values Entered on This Form
Values Computed from the Data Grid
Test Assumptions:
Independent Correlations
Dependent Correlations
Correlation r(x,y): .4
Correlation r(x,z): .6
Correlation(r(y,z): .7
Sample Size: 50
Percent Confidence Interval ? 95.0
Reset
Cancel
Continue

Fig. 4.9 Form for Comparison of Correlations

```
COMPARISON OF TWO CORRELATIONS

Correlation x with y =  0.400
Correlation x with z =  0.600
Correlation y with z =  0.700
Sample size = 50
Confidence Level Selected = 95.0
Difference r(x,y) - r(x,z) = -0.200
t test statistic = -2.214
Probability > |t| =  0.032
t value for significance =  2.012
```

Partial and Semi-Partial Correlations

Partial Correlation

One is often interested in knowing what the product-moment correlation would be between two variables if one or more related variables could be held constant. For example, in one of our previous examples, we may be curious to know what the correlation between achievements in learning French is with past achievement in learning English with intelligence held constant. In other words, if that proportion of variance shared by both French and English learning with IQ is removed, what is the remaining variance shared by English and French?

When one subtracts the contribution of a variable, say, X_3, from both variables of a correlation say, X_1 and X_2, we call the result the partial correlation of X_1 with X_2 partialling out X_3. Symbolically this is written as $r_{12.3}$ and may be computed by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \tag{3.70}$$

More than one variable may be partialled from two variables. For example, we may wish to know the correlation between English and French achievement partialling both IQ and previous Grade Point Average. A general formula for multiple partial correlation is given by

$$r_{12.34..k} = \frac{(1.0 - R^2_{y.34\,..\,k}) - (1.0 - R^2_{y.12\,..\,k})}{1.0 - R^2_{y.34\,..\,k}} \tag{3.71}$$

Semi-Partial Correlation

It is not necessary to partial out the variance of a third variable from both variables of a correlation. It may be the interest of the researcher to partial a third variable from only one of the other variables. For example, the researcher in our previous example may feel that intelligence should be left in the variance of the past English achievement which has occurred over a period of years but should be removed from the French achievement which is a much short learning experience. When the variance of a third variable is partialled from only one of the variables in a correlation, we call the result a semi-partial or part correlation. The symbol and calculation of the part correlation is

$$r_{1(2.3)} = \frac{r_{1,2} - r_{1,3}r_{2,3}}{\sqrt{(1.0 - r^2_{23})}} \tag{3.72}$$

where X_3 is partialled only from X_2 .

The squared multiple correlation coefficient R^2 may also be expressed in terms of semi_partial correlations. For example, we may write the equation

$$R^2_{y.1\,2\,..\,k} = r^2_{y.1} + r^2_{y(2.1)} + r^2_{y(3.12)} + .. + r^2_{y(k.12..k-1)} \tag{3.73}$$

In this formula, each semi-partial correlation reflects the proportion of variance contributed by a variable independent of previous variables already entered in the equation. However, the order of entry is important. Any given variable may explain a different proportion of variance of the independent variable when entered first, say, rather than last!

The semi-partial correlation of two variables in which the effects of K-1 other variables have been partialed from the second variable may be obtained by multiple regression. That is

$$r^2_{y(1.2\ 3\ ..\ k)} = R_{2y.1\ 2\ ..\ k} - R^2_{y.23..k} \qquad (3.74)$$

Autocorrelation

A large number of measurements are collected over a period of time. Stock prices, quantities sold, student enrollments, grade point averages, etc. may vary systematically across time periods. Variations may reflect trends which repeat by week, month or year. For example, a grocery item may sell at a fairly steady rate on Tuesday through Thursday but increase or decrease on Friday, Saturday, Sunday and Monday. If we were examining product sales variations for a product across the days of a year, we might calculate the correlation between units sold over consecutive days. The data might be recorded simply as a series such as "units sold" each day. The observations can be recorded across the columns of a grid or as a column of data in a grid. As an example, the grid might contain:

CASE/VAR	Day	Sold
Case 1	1	34
Case 2	2	26
Case 3	3	32
Case 4	4	39
Case 5	5	29
Case 6	6	14
...		
Case 216	6	15
Case 217	7	12

If we were to copy the data in the above "Sold" column into an adjacent column but starting with the Case 2 data, we would end up with:

CASE/VAR	Day	Sold	Sold2
1	34	26	
2	26	32	
3	32	39	
4	39	29	
5	29	14	
6	14	11	
...			
6	15	12	
7	12	-	

In other words, we repeat our original scores from Case 2 through case 217 in the second column but moved up one row. Of course, we now have one fewer case with complete data in the second column. We say that the second column of data "lags" the first column by 1. In a similar fashion we might create a third, fourth, fifth, etc. column representing lags of 2, 3, 4, 5, etc.. Creating lag variables 1 through 6 would result in variables starting with sales on days 1 through 7, that is, a week of sale data. If we obtain the product-moment correlations for these seven variables, we would have the correlations among Monday sales, Tuesday Sales, Wednesday Sales, etc. We note that the mean and variance are best estimated by the lag 0 (first column) data since it contains all of the observations (each lag loses one additional observation.) If the sales from day to day represent "noise" or simply random variations then we would expect the correlations to be close to zero. If, on the other hand, we see an systematic increase or decrease in sales between say, Monday and Tuesday, then we would observe a positive or negative correlation.

In addition to the inter-correlations among the lagged variables, we would likely want to plot the average sales for each. Of course, these averages may reflect simply random variation from day to day. We may want to "smooth" these averages to enhance our ability to discern possible trends. For example, we might want the average of day three to be a weighted average of that day plus the previous two day sales. This "moving average" would tend to smooth random peaks and valleys that occur from day to day.

It is also the case that an investigator may want to predict the sales for a particular day based on the previous sales history. For example, we may want to predict day 8 sales given the history of previous seven day sales.

Now let us look at an example of auto-correlation. We will use a file named strikes.tab. The file contains a column of values representing the number of strikes which occurred each month over a 30 month period. Select the auto-correlation procedure from the Correlations sub-menu of the Statistics main menu. Below is a representation of the form as completed to obtain auto-correlations, partial auto-correlations, and data smoothing using both moving average smoothing and polynomial regression smoothing:

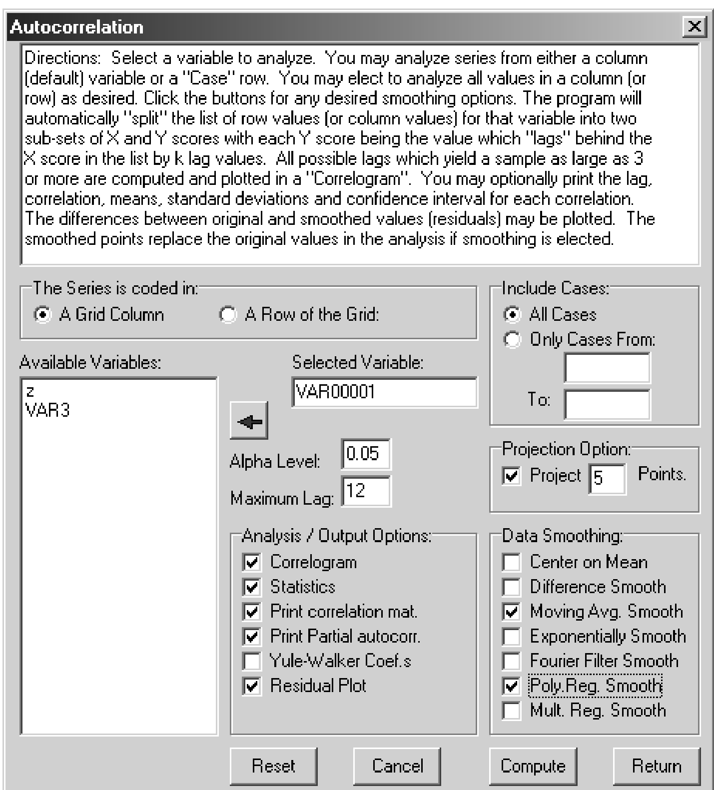


Fig. 4.10   The Autocorrelation Dialog

When we click the Compute button, we first obtain a dialog form for setting the parameters of our moving average. In that form we first enter the number of values to include in the average from both sides of the current average value. We selected 2. Be sure and press the Enter key after entering the order value. When you do, two theta values will appear in a list box. When you click on each of those thetas, you will see a default value appear in a text box. This is the weight to assign the leading and trailing averages (first or second in our example.) In our example we have accepted the default value for both thetas (simply press the Return key to accept the default or enter a value and press the Return key.) Now press the Apply button. When you do this, the weights for all of the values (the current mean and the 1, 2, … order means) are recalculated. You can then press the OK button to proceed with the process.

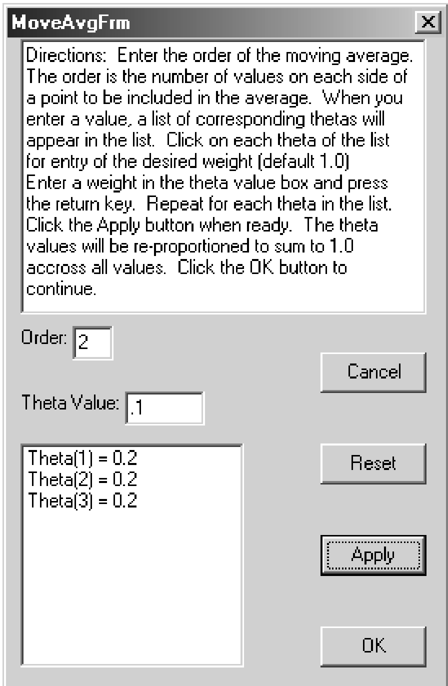


Fig. 4.11   The Moving Average Dialog

The procedure then plots the original (30) data points and their moving average smoothed values. Since we also asked for a projection of 5 points, they too are plotted. The plot should look like that shown below:

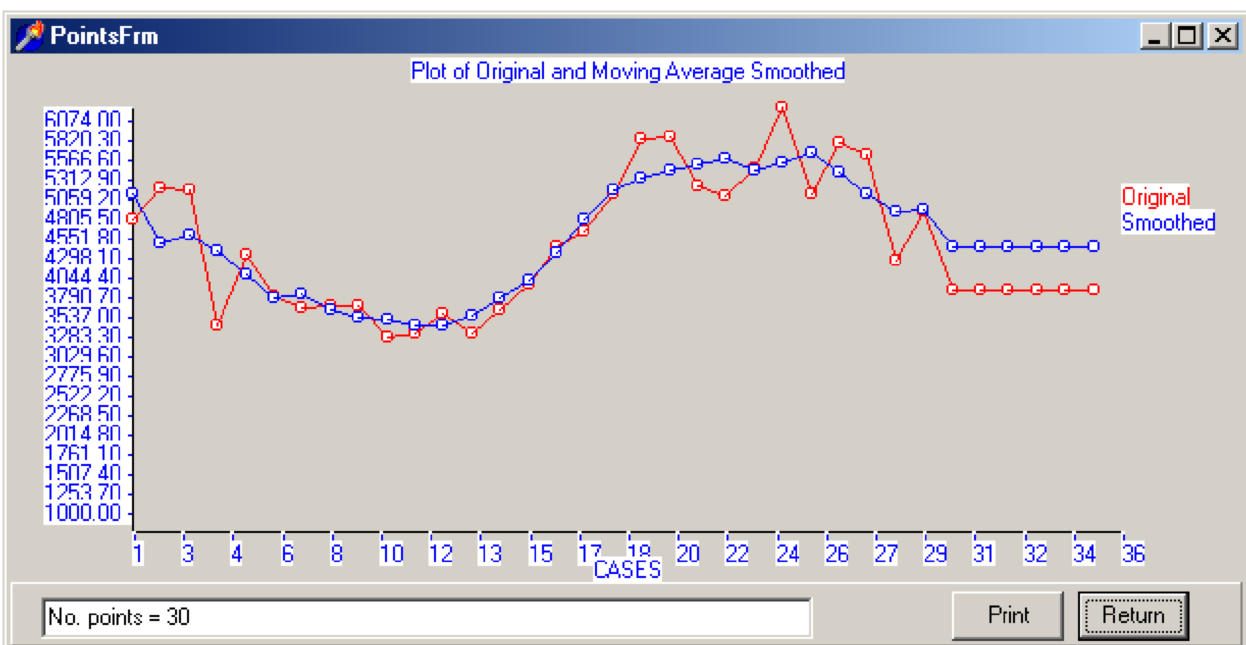


Fig. 4.12   Plot of Smoothed Points Using Moving Averages

We notice that there seems to be a “wave” type of trend with a half-cycle of about 15 months. When we press the Return button on the plot of points we next get the following:

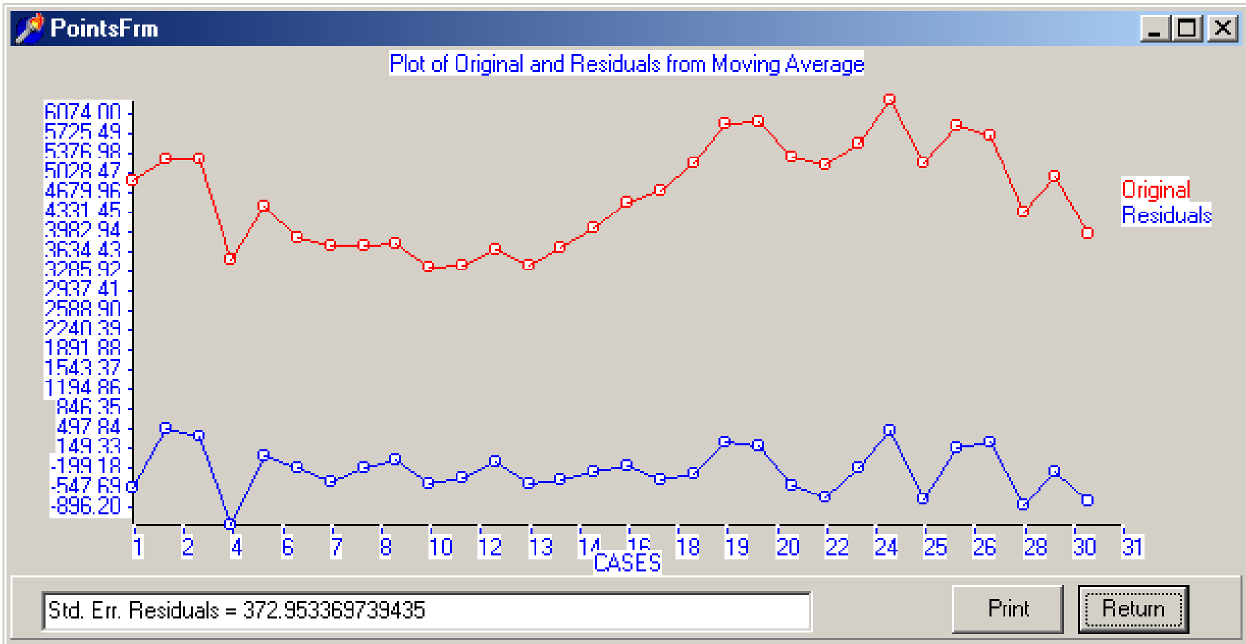


Fig. 4.13 Plot of Residuals Obtained Using Moving Averages

This plot shows the original points and the difference (residual) of the smoothed values from the original. At this point, the procedure replaces the original points with the smoothed values. Press the Return button and you next obtain the following:

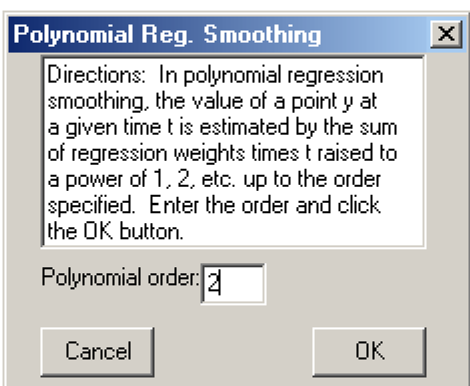


Fig. 4.14 Polynomial Regression Smoothing Form

This is the form for specifying our next smoothing choice, the polynomial regression smoothing. We have elected to use a polynomial value of 2 which will result in a model for a data point $Y_{t-1} = B * t^2 + C$ for each data point. Click the OK button to proceed. You then obtain the following result:

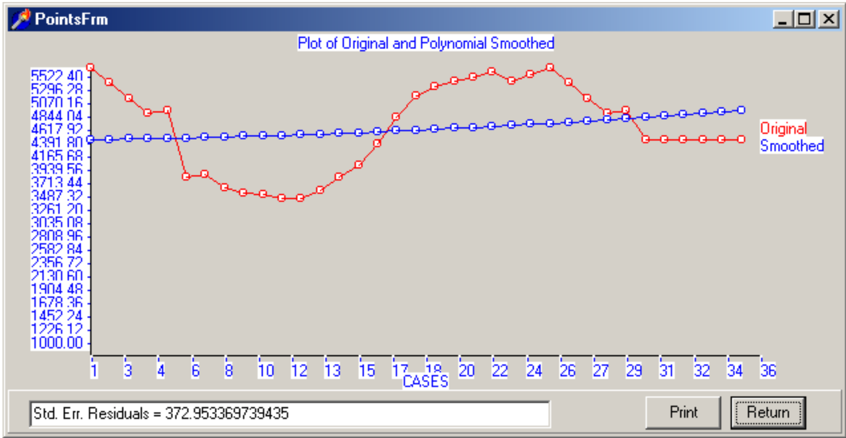


Fig. 4.15 Plot of Polynomial Smoothed Points

It appears that the use of the second order polynomial has "removed" the cyclic trend we saw in the previously smoothed data points. Click the return key to obtain the next output as shown below:

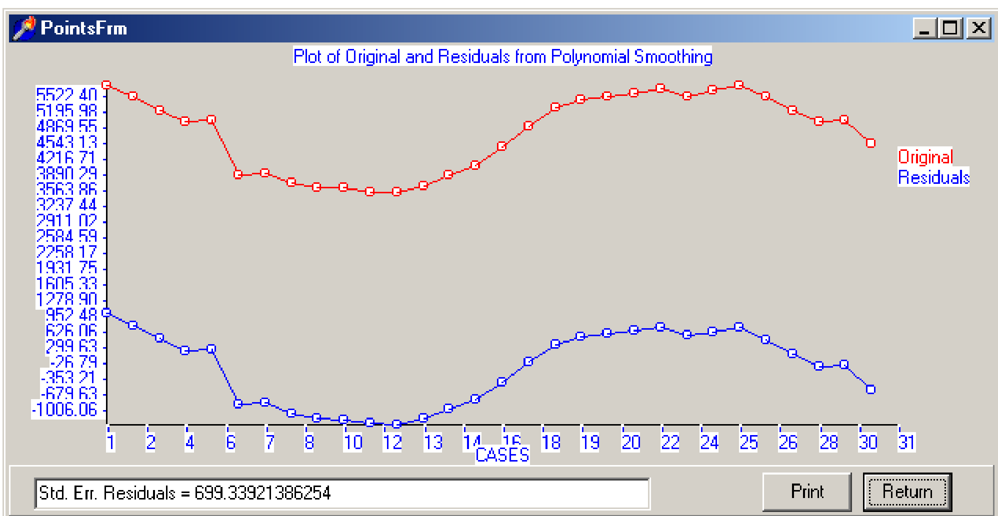


Fig. 4.16 Plot of Residuals from Polynomial Smoothing

This result shows the previously smoothed data points and the residuals obtained by subtracting the polynomial smoothed points from those previous points. Click the Return key again to see the next output shown below:

Overall mean = 4532.604, variance = 11487.241

Lag	Rxy	MeanX	MeanY	Std.Dev.X	Std.Dev.Y	Cases	LCL	UCL
0	1.0000	4532.6037	4532.6037	109.0108	109.0108	30	1.0000	1.0000
1	0.8979	4525.1922	4537.3814	102.9611	107.6964	29	0.7948	0.9507
2	0.7964	4517.9688	4542.3472	97.0795	106.2379	28	0.6116	0.8988
3	0.6958	4510.9335	4547.5011	91.3660	104.6337	27	0.4478	0.8444
4	0.5967	4504.0864	4552.8432	85.8206	102.8825	26	0.3012	0.7877
5	0.4996	4497.4274	4558.3734	80.4432	100.9829	25	0.1700	0.7287
6	0.4050	4490.9565	4564.0917	75.2340	98.9337	24	0.0524	0.6679
7	0.3134	4484.6738	4569.9982	70.1928	96.7340	23	-0.0528	0.6053
8	0.2252	4478.5792	4576.0928	65.3196	94.3825	22	-0.1470	0.5416
9	0.1410	4472.6727	4582.3755	60.6144	91.8784	21	-0.2310	0.4770
10	0.0611	4466.9544	4588.8464	56.0772	89.2207	20	-0.3059	0.4123
11	-0.0139	4461.4242	4595.5054	51.7079	86.4087	19	-0.3723	0.3481
12	-0.0836	4456.0821	4602.3525	47.5065	83.4415	18	-0.4309	0.2852

In the output above we are shown the auto-correlations obtained between the values at lag 0 and those at lags 1 through 12. The procedure limited the number of lags automatically to insure a sufficient number of cases upon which to base the correlations. You can see that the upper and lower 95% confidence limits increases as the number of cases decreases. Click the Return button on the output form to continue the process.

```
Matrix of Lagged Variable: VAR00001 with   30 valid cases.


Variables
                 Lag 0        Lag 1        Lag 2        Lag 3        Lag 4
     Lag 0       1.000        0.898        0.796        0.696        0.597
     Lag 1       0.898        1.000        0.898        0.796        0.696
     Lag 2       0.796        0.898        1.000        0.898        0.796
     Lag 3       0.696        0.796        0.898        1.000        0.898
     Lag 4       0.597        0.696        0.796        0.898        1.000
     Lag 5       0.500        0.597        0.696        0.796        0.898
     Lag 6       0.405        0.500        0.597        0.696        0.796
     Lag 7       0.313        0.405        0.500        0.597        0.696
     Lag 8       0.225        0.313        0.405        0.500        0.597
     Lag 9       0.141        0.225        0.313        0.405        0.500
    Lag 10       0.061        0.141        0.225        0.313        0.405
    Lag 11      -0.014        0.061        0.141        0.225        0.313
    Lag 12      -0.084       -0.014        0.061        0.141        0.225


Variables
                 Lag 5        Lag 6        Lag 7        Lag 8        Lag 9
     Lag 0       0.500        0.405        0.313        0.225        0.141
     Lag 1       0.597        0.500        0.405        0.313        0.225
     Lag 2       0.696        0.597        0.500        0.405        0.313
     Lag 3       0.796        0.696        0.597        0.500        0.405
     Lag 4       0.898        0.796        0.696        0.597        0.500
     Lag 5       1.000        0.898        0.796        0.696        0.597
     Lag 6       0.898        1.000        0.898        0.796        0.696
     Lag 7       0.796        0.898        1.000        0.898        0.796
     Lag 8       0.696        0.796        0.898        1.000        0.898
     Lag 9       0.597        0.696        0.796        0.898        1.000
    Lag 10       0.500        0.597        0.696        0.796        0.898
    Lag 11       0.405        0.500        0.597        0.696        0.796
    Lag 12       0.313        0.405        0.500        0.597        0.696


Variables
                Lag 10       Lag 11       Lag 12
     Lag 0       0.061       -0.014       -0.084
     Lag 1       0.141        0.061       -0.014
     Lag 2       0.225        0.141        0.061
     Lag 3       0.313        0.225        0.141
     Lag 4       0.405        0.313        0.225
     Lag 5       0.500        0.405        0.313
     Lag 6       0.597        0.500        0.405
     Lag 7       0.696        0.597        0.500
     Lag 8       0.796        0.696        0.597
     Lag 9       0.898        0.796        0.696
    Lag 10       1.000        0.898        0.796
    Lag 11       0.898        1.000        0.898
    Lag 12       0.796        0.898        1.000
```

The above data presents the inter-correlations among the 12 lag variables. Click the output form's Return button to obtain the next output:

```
Partial Correlation Coefficients with   30 valid cases.

Variables          Lag 0        Lag 1        Lag 2        Lag 3        Lag 4
                   1.000        0.898       -0.051       -0.051       -0.052

Variables          Lag 5        Lag 6        Lag 7        Lag 8        Lag 9
                  -0.052       -0.052       -0.052       -0.052       -0.051

Variables         Lag 10       Lag 11
                  -0.051       -0.051
```

The partial auto-correlation coefficients represent the correlation between lag 0 and each remaining lag with previous lag values partialled out. For example, for lag 2 the correlation of -0.051 represents the correlation between lag 0 and lag 2 with lag 1 effects removed. Since the original correlation was 0.796, removing the effect of lag 1 made a considerable impact. Again click the Return button on the output form. Next you should see the following results:

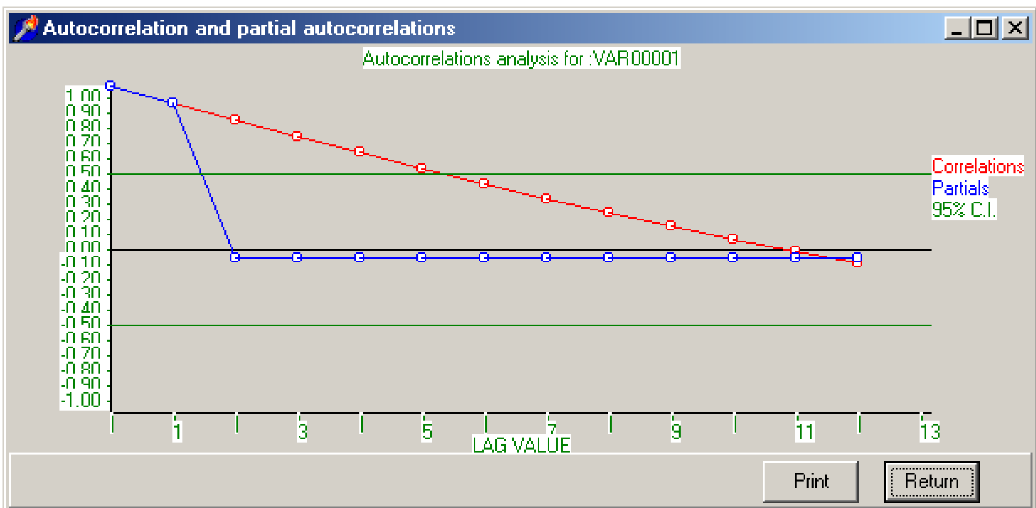


Fig. 4.17   Auto and Partial Autocorrelation Plot

This plot or "correlogram" shows the auto-correlations and partial auto-correlations obtained in the analysis. If only "noise" were present, the correlations would vary around zero. The presence of large values is indicative of trends in the data.

Series

Introduction

In many areas of research observations are taken periodically of the same object. For example, a medical researcher may take hourly blood pressure readings of a patient. An economist may record the price of a given stock each day for a long period. A retailer may record the number of units sold of a particular item on a daily basis. An industrialist may record the number of parts rejected each day over a period of time. In each of these cases, the researcher may be interested in identifying patterns in the fluctuation of the observations. For example, does a patient's systolic blood pressure systematically increase or decrease during visits by relatives? Do stock prices tend to vary systematically from month to month? Does the number of cans of tomato soup sold vary systematically across the days of the week or the months? Does the number of parts rejected in the assembly line vary systematically with the time of day or day of the week?

One approach often taken to discern patterns in repeated measurements is to simply plot the observed values across the time intervals on which the recording took place. This may work well to identify major patterns in the data. Sometimes however, factors which contribute to large systematic variations may "hide" other patterns that exist. A variety of methods have been developed to identify such patterns. For example, if the patterns are thought to potentially follow a sin wave pattern across time, a Fourier analysis may be used. This method takes a "signal" such as an electrical signal or a series of observations such as units sold each day and attempts to decompose the signal into fundamental frequencies. Knowing the frequencies allows the researcher to identify the "period" of the waves. Another method often employed involves examining the product-moment correlation between observations beginning at a specific "lag" period from each other. For example, the retailer may create an "X" variable beginning on a Monday and and "Y" variable beginning on the Monday four weeks later. The number of units sold are then recorded for each of these Mondays, Tuesdays, etc. If there is a systematic variation in the number of units sold over the weeks of this lag, the correlation will tend to be different from zero. If, on the other hand, there is only random variation, the correlation would be expected to be zero. In fact, the retailer may vary the lag period by 1 day, 2 days, 3 days, etc. for a large number of possible lag periods. He or she can then examine the correlations obtained for the various lags and where the correlations are larger, determine the pattern(s) that exist. One can also "co-vary out" the previous lag periods (i.e. get partial correlations) to identify whether or not more than one pattern may exist.

Once patterns of variability over time are identified, then observations at future time periods may be predicted with greater accuracy than one would obtain by simply using the average of all observations. The Auto-Regressive Imbedded Moving Average (ARIMA) method developed by Box and Jenkins is one such prediction tool. In that method, the relationship between a set of predictor observations and subsequent observations are optimized in a fashion similar to multiple regression or canonical correlation. When the interest is in predicting only a small number of future values, other methods may be employed such as multiple regression, moving average, etc.

The LazStats program provides the means for obtaining auto-correlations, partial auto-correlations, Fourier analysis, moving average analysis and other tools useful for time series analyses.

Calculating Correlations

Correlation

The Product Moment Correlation

It seems most living creatures observe relationships, perhaps as a survival instinct. We observe signs that the weather is changing and prepare ourselves for the winter season. We observe that when seat belts are worn in cars that the number of fatalities in car accidents decrease. We observe that students that do well in one subject tend to perform will in other subjects. This chapter explores the linear relationship between observed phenomena.

If we make systematic observations of several phenomena using some scales of measurement to record our observations, we can sometimes see the relationship between them by "plotting" the measurements for each pair of measures of the observations. As a hypothetical example, assume you are a commercial artist and produce sketches for advertisement campaigns. The time given to produce each sketch varies widely depending on deadlines established by your employer. Each sketch you produce is ranked by five marketing executives and an average ranking produced (rank 1 = best, rank 5 = poorest.) You suspect there is a relationship between time given (in minutes) and the average quality ranking obtained. You decide to collect some data and observe the following:

Average Rank (Y)	Minutes (X)
3.8	10
2.6	35
4.0	5
1.8	42
3.0	30
2.6	32
2.8	31

3.2	26
3.6	11
2.8	33

Using LazStats Descriptive menu's Plot X vs. Y procedure to plot these values yields the scatter-plot shown on the following page. Is there a relationship between the production time and average quality ratings?

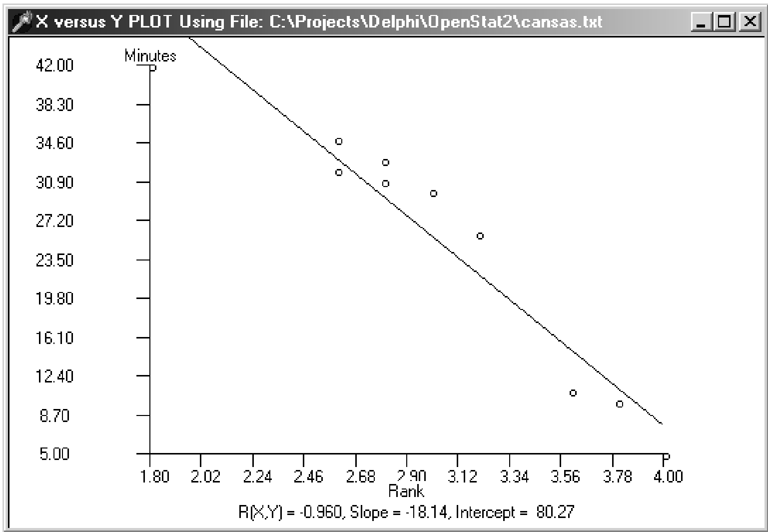


Fig. 4.18 X Versus Y Plot

LazStats contains a procedure for completing a z test for data like that presented above. Under the Statistics menu, move your mouse down to the Comparisons sub-menu, and then to the option entitled "One Sample Tests". When the form below displays, click on the Correlation button and enter the sample value .5, the population value .6, and the sample size 50. Change the confidence level to 90.0%.

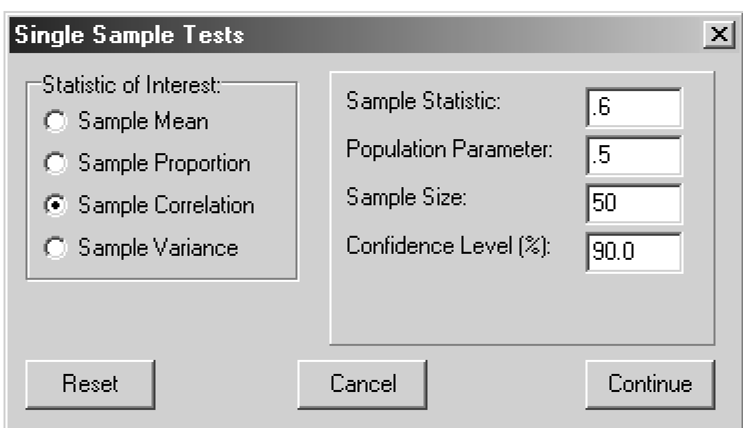


Fig. 4.19   SingleSample Tests Dialog Form

Shown below is the z-test for the above data:

```
ANALYSIS OF A SAMPLE CORRELATION

Sample Correlation =  0.600
Population Correlation =  0.500
```

```
Sample Size = 50
z Transform of sample correlation =  0.693
z Transform of population correlation =  0.549
Standard error of transform =  0.146
z test statistic =  0.986 with probability  0.838
z value required for rejection =  1.645
Confidence Interval for sample correlation = ( 0.425, 0.732)
```

Again, LazStats provides the computations for the difference between dependent correlations as shown in the Fig. below:

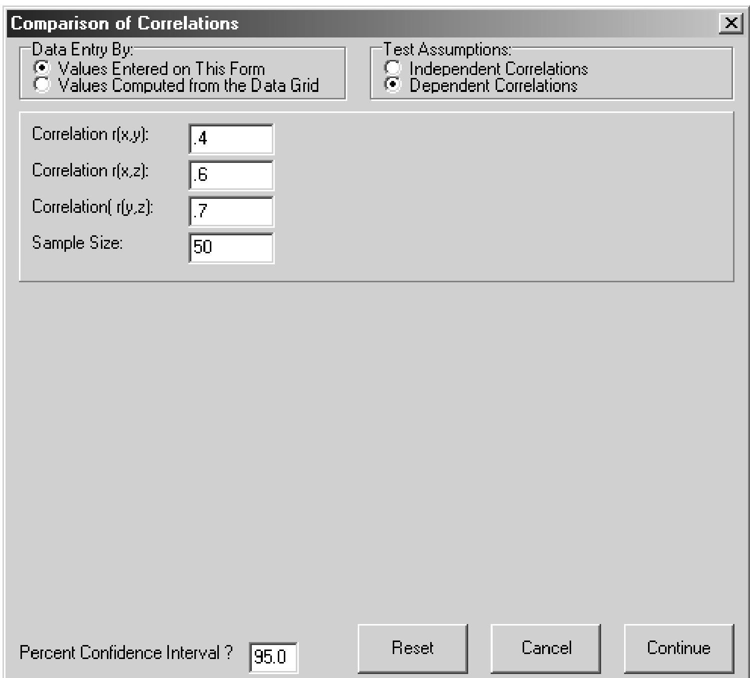


Fig. 4.20   Form for Comparison of Correlations

```
COMPARISON OF TWO CORRELATIONS

Correlation x with y =  0.400
Correlation x with z =  0.600
Correlation y with z =  0.700
Sample size = 50
Confidence Level Selected = 95.0
Difference r(x,y) - r(x,z) = -0.200
t test statistic = -2.214
Probability > |t| =  0.032
t value for significance =  2.012
```

Partial and Semi-Partial Correlations

Partial Correlation

One is often interested in knowing what the product-moment correlation would be between two variables if one or more related variables could be held constant. For example, in one of our previous analyses using the cansas.laz file, we may be curious to know what the correlation between jump height is with weight, waist and pulse with chins and situps held constant.

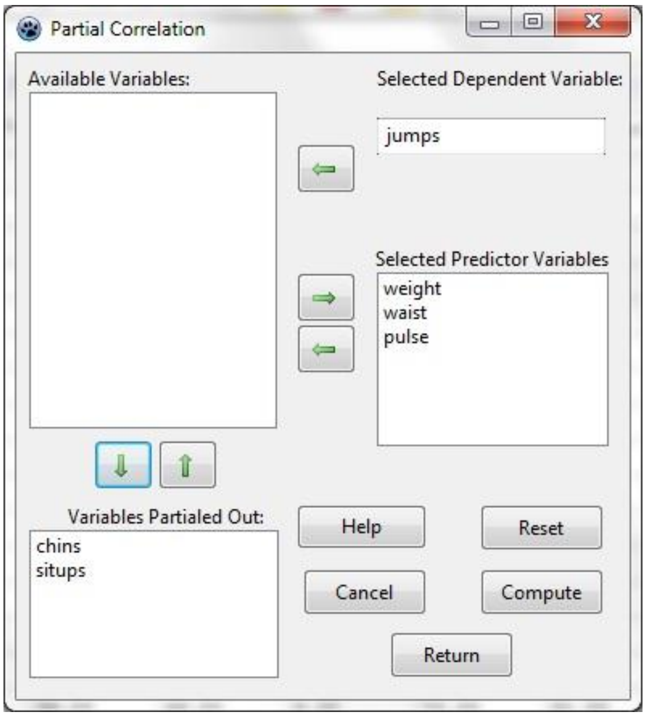


Fig. 4.21 Form for Partial and Semi-Partial Correlation

Partial and Semi-Partial Correlation Analysis

Dependent variable = jumps

Predictor Variables:
Variable 2 = weight
Variable 3 = waist
Variable 4 = pulse

Control Variables:
Variable 2 = chins
Variable 3 = situps

Higher order partialling at level = 3

Multiple partialling with 2 variables.

Squared Multiple Correlation with all variables = 0.636

Standardized Regression Coefficients:

weight = -0.588
waist = 0.982
pulse = -0.064
chins = 0.201
situps = 0.888

Squared Multiple Correlation with control variables = 0.450

Standardized Regression Coefficients:
chins = 0.058
situps = 0.629

Partial Correlation = 0.583

Semi-Partial Correlation = 0.432

F =   2.398 with probability = 0.1117, D.F.1 =   3 and D.F.2 = 14

Chapter 5. Multiple Regression

This chapter develops the theory and applications of Multiple Linear Regression Analysis. The multiple regression methods are frequently used (and misused.)   It also forms the heart of several other analytic methods including Path Analysis, Structural Equation Modeling and Factor Analysis.

The Linear Regression Equation

One of the major applications in statistics is the prediction of one or more characteristics of individuals on the basis of knowledge about related characteristics. For example, common-sense observation has taught most of us that the amount of time we practice learning something is somewhat predictive of how well we perform on that thing we are trying to master. Our bowling score tends to improve (up to a point) in relationship to the amount of time we spend practicing bowling. In the social sciences however, we are often interested in predicting less obvious outcomes. For example, we may be interested in predicting how much a person might be expected to use a computer on the basis of a possible relationship between computer usage and other characteristics such as anxiety in using machines, mathematics aptitude, spatial visualization skills, etc. Often we have not even observed the relationships but instead must simply hypothesize that a relationship exists. In addition, we must hypothesize or assume the type of relationship between our variables of interest. Is the relationship a linear one? Is it a curvilinear one?

Multiple regression analysis is a method for examining the relationship between one continuous variable of interest (the dependent or criterion variable) and one or more independent (predictor) variables. Typically we assume a linear relationship of the type

$$Y_i = B_1X_{i1} + B_2X_{i2} + ... + B_kX_{ik} + B_0 + E_i \qquad (5.1)$$

where

Y_i is the score obtained for individual i on the dependent variable,

X_{i1} ... X_{ik} are scores obtained on k independent variables,

B_1 ... B_k are weights (regression coefficients) of the k independent variables which maximize the relationship with the Y scores,

B_0 is a constant (intercept) and E_i is the error for individual i.

In the above equation, the error score E_i reflects the difference between the subject's actual score Yi and the score which is predicted on the basis of the weighted combination of the independent variables. That is,

$$Y'_i - Y_i = E_i . \qquad (5.2)$$

where Y'_i is predicted from

$$Y'i = B_1X_{i1} + B_2X_{i2} + ... + B_kX_{ik} + B_0 \qquad (5.3)$$

In addition to assuming the above general linear model relating the Y scores to the X scores, we usually assume that the E_i scores are normally distributed.

When we complete a multiple regression analysis, we typically draw a sample from a population of subjects and observe the Y and X scores for the subjects of that sample. We use that sample data to estimate the weights (B's) that will permit us the "best" prediction of Y scores for other individuals in the population for which we only have knowledge of their X scores. For example, assume we are interested in predicting the scores that individuals make on a paper and pencil final examination test in a statistics course in graduate college. We might hypothesize that students who, in the past, have achieved higher grade point averages as undergraduates would likely do better on a statistics test. We might also suspect that students with higher mathematics aptitudes as measured by the mathematics score on the Graduate Record Examination would do better than students with lower scores. If undergraduate GPA and GRE-Math combined are highly related to achievement on a graduate statistics grade, we could use those two variables as predictors of success on the statistics test. Note that in this example, the GRE and undergraduate GPA are obtained for individuals quite some time before they even enroll in the statistics course! To find that weighted combination of GRE and GPA scores which "best" predicts the graduate statistics grades of students, we must observe the actual grades obtained by a sample of students that take the statistics course.

Notice that in our linear prediction model, we are going to obtain, for each individual, a single predictor score that is a weighted combination of independent variable scores. We could, in other words, write our prediction equation as

$$Y'_i = C_i + B_0 \tag{5.4}$$

where

$$C_i = \sum_{j=1}^{k} B_i X_I \tag{5.5}$$

You may recognize that equation (3) above is a simple linear equation. The product-moment correlation between Yi and Ci in equation (3) is an index of the degree to which the dependent and composite score are linearly related. In a previous chapter we expressed this relationship with r_{xy} and the proportion of variance shared as r^2_{xy}. When x is replace by a weighted composite score C, we differentiate from the simple product-moment correlation by use of a capital r, that is $R_{y.1,2,..,k}$ with the subscripts after the period indicating the k independent variables. The proportion of variance of the Y scores that is predicted by weighted composite of X scores is, similarly, $R^2_{y.1,2,..,k}$.

We previously learned that, for one independent variable, the "best" weight (B) could be obtained from

$$B = r_{xy} S_y / S_x . \tag{5.6}$$

We did not, however, demonstrate exactly what was meant by the best fitting line or best B. We need to learn how to calculate the values of B when there is more than one independent variable and to interpret those weights.

In the situation of one dependent and one independent variable, the regression line is said to be the "best" fitting line when the squared distance of each observed Y score summed across all Y scores is a minimum. The Fig. on the following page illustrates the "best fitting" line for the pairs of x and y scores observed for five subjects. The line represents, of course, the equation

$$Y'_i = BX_i + B_0 \tag{5.7}$$

That is, the predicted Y value for any value of X. (See chapter III to review how to obtain B and B_0 .) Since we have defined error (E_i) as the difference between the observe dependent variable score (Y_i) and the predicted score, then our "best fitting" line is drawn such that

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - Y'_i)^2 \text{ is a minimum.} \tag{5.8}$$

We can substitute our definition of Y'_i from equation (4.7) above in equation (4.8) above and obtain

$$G = \sum_{i=1}^{n} [Yi - (BX_i + B_0)]^2 = \text{a minimum} \tag{5.9}$$

Expanding equation (5.9) yields

$$G = \sum_{i=1}^{n} Y_i^2 + \sum_{i=1}^{n} (BX_i + B_0)^2 - 2 \sum_{i=1}^{n} Y_i(BX_i + B_0) \tag{5.10}$$

$$= \sum_{i=1}^{n} Y_i^2 + \sum_{i=1}^{n}(B^2X_i^2 + B_0^2 + 2B_0BX_i) - 2B \sum_{i=1}^{n} Y_iX_i - 2B_0 \sum_{i=1}^{n} Y_i$$

or

$$G = \sum_{i=1}^{n} Y_i^2 + B^2 \sum_{i=1}^{n} X_i^2 + nB_0^2 + 2B_0B \sum_{i=1}^{n} X_i - 2B \sum_{i=1}^{n} Y_iX_i - 2B_0 \sum_{i=1}^{n} Y_i \tag{5.11}$$

$$= \text{ a minimum.}$$

Notice that the function G is affected by two unknowns, B_0 and B. There is one pair of these values which makes G a minimum value _ any other pair would cause G (the sum of squared errors) to be larger. But how do we determine B and B_0 that guarantees, for any observed sample of data, a minimum G? To answer this question requires we learn a little bit about minimizing a function. We will introduce some very elementary concepts of Calculus in order to solve for values of B and B_0 that minimize the sum of square errors.

Least Squares Calculus

Definitions:

Definition 1: A function (f) is a correspondence between the objects of one class and those of another which pairs each member of the first class with one and only one member of the second class. We have several ways of specifying functions, for example, we might provide a complete cataloging of all the associated pairs, e.g.

Class 1 (x)	1	2	3	4	5
class 2 f(x)	3	5	7	9	11

where class 2 values are a function of class 1 values.

Another way of specifying a function is by means of a set of ordered pairs, e.g.

{ (1,3), (2,5), (3,7), (4,9), (5,11) }

We may also use a map or graph such as

```
15 |
14 |
13 |
12 |
11 |                  *
10 |
 9 |               *
 8 |
 7 |            *
 6 |
 5 |         *
 4 |
 3 |      *
 2 |
 1 |
 0 |
______________________________________
   0  1  2  3  4  5  6
```

Fig. 5.1 A Simple Function Map

Finally, we may use a mathematical formula:

$f(x) = 2X + 1$ where $X = 1,2,3,4,5$

Definition 2: Given a specific member of the first class, say X, the member of the second class corresponding to this first class member, designated by f(X), is said to be the value of the function at X.

Definition 3: The set of all objects of the first class is said to be the domain of the function. The set of all objects of the second class is the range of the function f(X).

In our previous example under definition 1, the domain is the set of numbers (1,2,3,4,5) and the range is (3,5,7,9,11). As another example, let X = any real number from 1 to 5 and let $f(X) = 2X + 1$. Then the domain is

$\{ X : 1 \leq X \leq 5 \}$ and the range is

$\{ f(X) : 3 \leq f(X) \leq 11 \}$.

Definition 4: The classes of objects or numbers referred to in the previous definitions are sometimes called variables. The first class is called the independent variable and the second class is called the dependent variable.

Definition 5: A quantity which retains a fixed value throughout the course of a discussion is called a constant. Some constants retain the same values in all discussions, e.g.

$\pi = c/d = 3.1416...$, and

$e =$ limit as $x \rightarrow \infty$ of $(1 + X)^{1/X} = 2.7183...$.

Other constants retain the save values in a given discussion but may vary from one discussion to another. For example, consider the function

$$f(X) = bX + a. \tag{5.12}$$

In the example under definition 1, b = 2 and a = 1. If b = -2 and a = 3 then the function becomes

f(x) = -2X + 3 .

If X is continuous or an infinite set, complete listing of the numbers is impossible but a map or formula may be used.   Now consider

X	1	2	2	3
f(X)	3	5	7	4

This is not a legitimate function as by definition there is not a one and only one correspondence of members.

Sometimes the domain is itself a set of ordered pairs or the sub-set of a plane. For example

```
          f(X,Y)
               3   |
                   |  :
               2   |  :
                   |  :
               1   |  :
                   |  :
               ___|_:________________________
                  /   :    1      2      3
             1   /    :  .      .
                /     :.      .
           2   /      :      .
              /
         3   /
            /
```

Fig. 5.2 A Function Map in Three Dimensions

The domain of { (X,Y) : 0 $\leq$ X $\leq$ 2 & 0 $\leq$ Y $\leq$ 2 }

f(X,Y) = 2X + Y + 1

Range of { 1 $\leq$ f(X,Y) $\leq$ 7 }

Finding A Change in Y Given a Change in X For Y=f(X)

It is often convenient to use Y or some other letter as an abbreviation for f(X).

Definition 6: ΔX represents the amount of change in the value of X and ΔY represents the corresponding amount of change in the value of Y = f(X). ΔX and ΔY are commonly called increments of change or simply increments. For example, consider Y = f(X) = X^2 where:

Domain is { X : $-\infty$ < X < $+\infty$ }

Now let X = 5. Then Y = f(X) = 25. Now let ΔX = +2. Then

Y = +24. Or let ΔX = -2 then Y = -16. Finally,

let ΔX = 1/2 then Y = 5.25.

Trying a different starting point X = 3 and using the same values of X we would get:

if $X = 3$
and $\Delta X = +2$ then $Y = +16$
$\Delta X = -2$ then $Y = -8$
$\Delta X = .5$ then $Y = 3.25$

It is impractical to determine the increment in Y for an increment in X in the above manner for the general function $Y = f(X) = X^2$. A more general solution for Y is obtained by writing

$$Y + \Delta Y = f(X + \Delta X) = (X + \Delta X)^2$$

or, solving for Y by subtracting Y from both sides gives

$$Y = (X + \Delta X)^2 - Y \tag{5.13}$$

or $Y = X^2 + \Delta X^2 + 2X \Delta X - Y$

or $Y = X^2 + \Delta X^2 + 2X \Delta X - X^2$

or $Y = 2X \Delta X + \Delta X^2$

Now using this formula:

If X = 5 and $\Delta X = 2$ then Y = +24 or if X = 5 and $\Delta X = -2$ then Y = -16. These values are the same as we found by our previous calculations!

Relative Change in Y for a Change in X

We may express the relative change in a function with respect to a change in X as the ratio

$$\frac{\Delta Y}{\Delta X}$$

For the function $Y = f(X) = X^2$ we found that $Y = 2X \Delta X + \Delta X^2$

Dividing both sides by ΔX we then obtain

$$\frac{\Delta Y}{\Delta X} = 2X + \Delta X \tag{5.14}$$

For example, when X = 5 and X = +2, the relative change is

$$\frac{\Delta Y}{\Delta X} = \frac{24}{2} = 2(5) + 2 = 12$$

The Concept of a Derivative

We may ask what is the limiting value of the above ratio (ΔY / ΔX) of relative change is as the increment in X (ΔX) approaches 0 ($\Delta X \rightarrow 0$). We use the symbol

$$\frac{dY}{dX} \quad \text{to represent this limit.}$$

We note that for the function $Y = X^2$, the relative change was

$$\frac{\Delta Y}{\Delta X} = 2X + \Delta X . \tag{5.15}$$

If ΔX approaches 0 then the limit is

$$\frac{dY}{dX} = 2X .$$

Definition 7: The derivative of a function is the limit of a ratio of the increment of change of the function to the increment of the independent variable when the latter increment approaches 0 as a limit. Symbolically,

$$\frac{dY}{dX} = \lim_{\Delta X \rightarrow 0} \frac{\Delta Y}{\Delta X} = \lim_{\Delta X \rightarrow 0} \frac{f(X + \Delta X)-f(X)}{\Delta X}$$

Since $Y + \Delta Y = f(X + \Delta X)$ and $Y = f(X)$ then

$\Delta Y = f(X + \Delta X) - f(X)$ and the ratio

$$\frac{\Delta Y}{\Delta X} = \frac{f(X + \Delta X) - f(X)}{\Delta X} \tag{5.16}$$

EXAMPLE: $Y = X^2$ $dY/dX = ?$

$$dY/dX = \lim_{\Delta X \rightarrow 0} \frac{f(X + \Delta X) - f(X)}{\Delta X}$$

$$= \lim_{\Delta X \rightarrow 0} \frac{X^2 + \Delta X^2 + 2X\,\Delta X - X^2}{\Delta X}$$

$$= \lim_{\Delta X \rightarrow 0} \Delta X + 2X$$

or $$\frac{dY}{dX} = 2X$$

Some Rules for Differentiating Polynomials

Rule 1. If $Y = CX^n$, where n is an integer, then

$$\frac{dY}{dX} = nCX^{n-1} \tag{5.17}$$

For example, let C = 7 and n = 4 then $Y = 7X^4$.

$$\frac{dY}{dX} = (4)(7)X^3$$

Proof:

$$\frac{dY}{dX} = \lim_{\Delta X \to 0} \frac{C(X + \Delta X)^n - CX^n}{\Delta X}$$

$$\text{since } (a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$$

then

$$\frac{dY}{dX} = \lim_{\Delta X \to 0} [\, C\binom{n}{n}X^n \Delta X^{n-n} + C\binom{n}{n-1}X^{n-1} \Delta X^1$$

$$+ C\binom{b}{n-2}X^{n-2} \Delta X^2 + ...+ C\binom{b}{0}X^0 \Delta X^n$$

$$- CX^n] / \Delta X$$

$$= \lim_{\Delta X \to 0} [\, CX^n + C^nX^{n-1} \Delta X + C\frac{n(n\_1)}{2} X^{n-2} \Delta X^2$$

$$+ ...+ C \Delta X^n - CX^n \,] / \Delta X$$

$$= \lim_{\Delta X \to 0} CnX^{n-1} + \frac{n(n\_1)}{2} X^{n-2} \Delta X +...+ C \Delta X^{n-1}$$

or

$$\frac{dY}{dX} = CnX^{n-1} \qquad \text{(End of Proof)}$$

Rule 1.a If Y = CX then dY/dX = C (5.18)

since by Rule 1 $dY/dX = (1)CX^0 = C$

Rule 1.b If Y = C then dY/dX = 0

Note that dY/dX of CX^0 is $(0)CX^{-1} = 0$.

Rule 2. If Y = U + V - W where U, V and W are functions of X, then:

$$\frac{dY}{dX} = \frac{dU}{dX} + \frac{dV}{dX} - \frac{dW}{dX} \tag{5.19}$$

Example: Consider $Y = 4X^2 - 4X + 1$

Let $U = f(X) = 4X^2$ and

$V = f(X) = -4X$ and

$W = f(X) = 1$.

Applying Rules 1 and 2 we have

$$\frac{dY}{dX} = 8X - 4$$

Rule 3. If $U = V^n$ where V is a function of X then

$$\frac{dY}{dX} = nV^{n-1} \frac{dU}{dX} \tag{5.20}$$

Example: Consider $Y = (2X - 1)^2$

Let V = (2X - 1) and n = 2

Then

$$\frac{dY}{dX} = 2(2X - 1)(2) = 8X - 4$$

Another Example. Let $Y = \sum_{i=1}^{N} (3X + W_i)^2$

where W_i and N are variable constants,

that is, in one discussion $N_1 = 3$ and

$W_1 = 2$ or $W_2 = 4$ and $W_3 = 3$.

If, for example, X = 0, $Y = 2^2 + 4^2 + 3^2 = 29$

or, if X = 1 then $Y = 5^2 + 7^2 + 6^2 = 110$

Now we ask, dY / dX = ?

Solution:

$$\frac{dY}{dX} = \sum_{i=1}^{N} 2(3X + Wi)(3)$$

because $Y = (3X + W_1)^2 + (3X + W_2)^2 + (3X + W3)^2$

and applying Rules 2 and 3 we get:

$$\frac{dY}{dX} = 6 \sum_{i=1}^{N} (3X + W_i)$$

$$= 6 \sum_{i=1}^{N} 3X + 6 \sum_{i=1}^{N} W_i$$

$$= 6[N(3X)] + 6 \sum_{i=1}^{N} W_i$$

or

$$\frac{dY}{dX} = 18NX + 6 \sum_{i=1}^{N} W_i$$

Geometric Interpretation of a Derivative

The Fig. below presents a graphical representation of a function Y = f(X) (the curved line). Two points on the function are denoted as P(X,Y) and P(X + X,Y + Y). A straight line, a secant line, is drawn through the two points. Notice that if X becomes smaller (and therefore the corresponding Y becomes smaller) that the secant line approaches a tangent line at the point P(X,Y). We review:

$f(X) = Y$

$f(X + \Delta X) = (Y + \Delta Y)$ or $f(X + \Delta X) - f(X) = Y$

and $\dfrac{f(X + \Delta X) - f(X)}{\Delta X} = \dfrac{\Delta Y}{\Delta X}$

Note that $\Delta Y / \Delta X$ give rise over run or the slope of the of the secant line through two points on the function. Now if $X \rightarrow 0$, then P' approaches P and the secant line approaches a tangent at the point P. Therefore the dY / dX is the slope of the tangent at P or X.

We will now use the derivative in determining maximum points on a function.

Finding the Value of X for Which f(X) is Least.

Given the function $f(X) = Y = X^2 - 3X$ where $-\infty < X < +\infty$ we may present the function as in Fig. XII.2 below.

For the function, we may obtain some values of Y corresponding to a selected set of X values:

X	-2	-1	0	+1	+2	+3	+4	+5
Y	10	4	0	-2	-2	0	+4	+10

Then the derivative

$\dfrac{dY}{---}$ = 2X - 3 which is the slope of the tangent at any point X.

dX

Setting the slope (dY / dX) equal to zero we obtain the minimum value of X, that is,

$0 = 2X - 3$ and therefore $X = 1.5$ for a minimum Y value.

Another Example of a Minimum

Given a collection of score values X

{ X | 16, 8, 10, 4, 12 }

we ask for what value of A is f(A) a minimum if

$$f(A) = \sum_{i=1}^{5} (Xi - A)^2 \ ?$$

First, examine the f(A) for various values of A, for example:

if $A = 5$ then $f(A) = 11^2 + 3^2 + 5^2 + (-1)^2 + 72$
if $A = 7$ then $f(A) = 9^2 + 1^2 + 3^2 + (-3)^2 + 52$
if $A = 8$ then $f(A) = 8^2 + 0^2 + 2^2 + (-4)^2 + 42$
etc.

A plot of the function f(A) is presented below for the values

A	5	7	8	9	11	13	15
f(A)	205	125	100	85	85	125	205

```
f(A)
------
210|                     •                                       •
200|
190|
180|
170|
160|
150|
140|
130|                             •               •       •
120|
110|
100|                                 •
 90|                                     •
------- (minimum)
     0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
                                   A
```

Fig. 5.3 The Minimum of a Function Derivative

The derivative of the f(A) with respect to A is

$$\frac{d\,f(A)}{dA} = \sum_{i=1}^{5} 2(X_i - A)(-1) \qquad (5.21)$$

and to obtain the minimum slope point we obtain

$$0 = \sum_{i=1}^{5} -2(X_i - A) = \sum_{i=1}^{5} X_i - 5A \qquad (5.22)$$

or $A = \sum_{i=1}^{5} X_i / 5$

Therefore A = (16 + 8 + 10 + 4 + 12) / 5 = 10

and f(A) = 80 is a minimum.

A Generalization of the Last Example

We will use derivation to prove that given any collection of X values X1, X2, ..., Xi, ..., XN that

$$Y = \sum_{i=1}^{N} (Xi - A)^2 \text{ is least when } A = \bar{X}.$$

As before, the derivative of Y with respect to A is

$$\frac{dY}{dA} = \sum_{i=1}^{N} 2(X_i - A)(-1) = -2 \sum_{i=1}^{N} (X_i) - 2NA(-1)$$

Therefore if we set the derivative to zero we obtain

$$0 = -2 \sum_{i=1}^{N} X_i + 2NA$$

or $0 = - \sum_{i=1}^{N} X_i + NA$

then $A = \sum_{i=1}^{N} Xi / N$ (5.23)

which by definition is $\bar{X}$.

Partial Derivatives

Given a function in two independent variables:

$Y = f(X,Z)$

we may create a graph as shown if Fig. XII.4 below. Y, the function, is shown as the vertical axis and X and Z are shown as horizontal axis. Note the line in the Fig. which represents the map of f(X,Z) when one considers only one value of Z.

When we study functions of this type with one variable treated as a constant, the derivative of the function is called a partial derivative.

Suppose the function has a minimum and that it occurs at X = A and Z = B, that is, f(A,B) is a minimum value of Y. We may obtain the derivative of Y = f(A,Z), that is, treat Z as a constant. This would be the partial derivative δY/δZ and may be set equal to 0 to get the minimum at B. Of course, we don't know A. Likewise, Y = f(X,B) and δY/δX set equal to 0 will give A. Here we don't know B.

We can however, by simultaneous equations, where A and B are set to 0, find a minimum of X and Z to give the Y minimum.

For example, let $Y = f(X,Z) = X^2 + XZ + Z^2 - 6X + 2$.

Then

$$\frac{\delta Y}{\delta X} = 2X + Z - 6 = 0 \qquad (1)$$

and

$$\frac{\delta Y}{\delta Z} = X + 2Z = 0 \qquad (2)$$

or $X = -2Z$ for equation (2) and substituting in (1) gives

$-4Z + Z = 6$ or $Z = -2$

and therefore X = +4. These values of Z and X are the values of A and B to produce a minimum for Y = f(A,B).

Least Squares Regression for Two or More Independent Variables

In this section we want to use the concepts of partial derivation to obtain solutions to the B values in

$$Y'_i = B_1X_{i,1} + B_2X_{i,2} + B_0 \qquad (5.24)$$

such that the sum of $(Y - Y')^2$ is a minimum.

As an example, assume we have a situation in which values of Y_i represent Grade Point Average (GPA) score of subject (i) in his or her freshman year at college. Assume that the $X_{i,1}$ score is the high school GPA and that the $X_{i,2}$ is an aptitude test score. Our population of subjects may be "decomposed" into sub-populations of Y scores that correspond to given values of X_1 and X_2. Fig. XII.5 depicts the distributions of Y scores for combinations of X_1 and X_2 scores. We will assume:

(1) the experience pool of the available data is a random sample of (Y, X_1 and X_2) triplets from a universe of such triplets,

(2) the universe is capable of decomposition into sub-universes of triplets have like X_1 and X_2 values but differing in Y values,

(3) the Y means for the sub-universes fall on a plane, that is,

$$\mu_{Y,12} = \beta_1 X_1 + \beta_2 X_2 + \beta_0 \tag{5.25}$$

Now we use the data to estimate β_1 , β_2 and β_0 by finding those values of B_1 and B_2 and B_0 in:

$$Y' = B_1 X_1 + B_2 X_2 + B_0 \tag{5.26}$$

which minimize

$$G = \sum_{i=1}^{N} (Y_i - Y'_i)^2$$

or $$G = \sum_{i=1}^{N} [\, Y_i - (B_1 X_{i,1} + B_2 X_{i,2} + B_0)]^2 \tag{5.27}$$

The steps to our solution are:

1. Find the partial derivatives and equate them to 0.

$$\frac{\delta G}{\delta B_1} = 2 \sum_{i=1}^{N} [Y_i - (B_1 X_{i,1} + B_2 X_{i,2} + B_0)](-X_{i,1})$$

$$\frac{\delta G}{\delta B_2} = 2 \sum_{i=1}^{N} [Y_i - (B_1 X_{i,1} + B_2 X_{i,2} + B_0)](-X_{i,2})$$

$$\frac{\delta G}{\delta B_0} = 2 \sum_{i=1}^{N} [Y_i - (B_1 X_{i,1} + B_2 X_{i,2} + B_0)](-1)$$

Now equating to 0 and simplifying results in the following three "normal" equations:

$$\sum_{i=1}^{N} Y_i X_{i,1} = B_1 \sum_{i=1}^{N} X^2_{i,1} + B2 \sum_{i=1}^{N} X_{i,1} X_{i,2} + B_0 \sum_{i=1}^{N} X_{i,1} \tag{5.28}$$

$$\sum_{i=1}^{N} Y_i X_{i,2} = B_1 \sum_{i=1}^{N} X_{i,1} X_{i,2} + B_2 \sum_{i=1}^{N} X^2_{i,2} + B_0 \sum_{i=1}^{N} X_{i,2} \tag{5.29}$$

$$\sum_{i=1}^{N} Y_i = B_1 \sum_{i=1}^{N} X_{i,1} + B_2 \sum_{i=1}^{N} X_{i,2} + NB_0 \tag{5.30}$$

2. Use the data to obtain the various sums, sums of squared values, and sums of products needed. Substitute them in the above equations (4.28), (4.29) and (4.30) and solve the equations simultaneously for B_1, B_2 and B_0.

3. Substitute obtained values of B_1, B_2 and B_0 into equation 4.27 to get the regression equation.

4. If an index of accuracy of prediction is desired, calculate

$$\sum_{i=1}^{N} y'^2_i \text{ and obtain } R^2_{y.12} = \frac{\sum_{i=1}^{N} y'^2_i}{\sum_{i=1}^{N} y^2_i} \tag{5.31}$$

where the y'_i and y_i scores are deviations from the mean Y value.

Matrix Form for Normal Equations Using Raw Scores

Equations (4), (5) and (6) above may be written more conveniently in matrix form as:

$$\left[\sum_{i=1}^{N} Y_i X_{i,1} \;\; \sum_{i=1}^{N} Y_i X_{i,2} \;\; \sum_{i=1}^{N} Yi \right] =$$

$$[B_1\ B_2\ B_0] \begin{vmatrix} \sum_{i=1}^{N} X^2_{i,1} & \sum_{i=1}^{N} X_{i,1}X_{i,2} & \sum_{i=1}^{N} X_{i,1} \\ \sum_{i=1}^{N} X_{i,1}X_{i,2} & \sum_{i=1}^{N} X^2_{i,2} & \sum_{i=1}^{N} X_{i,2} \\ \sum_{i=1}^{N} X_{i,1} & \sum_{i=1}^{N} X_{i,2} & N \end{vmatrix}$$

or $[\, Y'X \,]_{1x\,(K+1)} = [\, B \,]'_{1\,x\,(K+1)}\ [\, X'X \,]_{(K+1)(K+1)}$

and leaving off the sizes of the matrices gives simply

$$[\, Y'X \,] = [\, B \,]'\ [\, X'X] .$$

If we post-multiply both sides of this equation by $[X'X]^{-1}$ we obtain

$$[\, Y'X \,]\ [X'X]\text{-}1 = [\, B \,]' \tag{5.32}$$

We note that B_0 may also be obtained from

$$B_0 = \overline{Y} - (B_1\overline{X}_1 + \ldots + B_k\overline{X}_k) \tag{5.33}$$

or in matrix notation

$$B_0 = \overline{Y} - [B]'[\overline{X}] \tag{5.34}$$

where $[\overline{X}] = (1/N)\ [X]$

Matrix Form for Normal Equations Using Deviation Scores

The prediction (regression) equation above may be written in deviation score form as

$$y' = B_1 x_{i,1} + B_2 x_{i,2} \tag{5.35}$$

and solve for $G = \sum_{i=1}^{N} (y_i - y'_i)^2$ as a minimum.

In deviation score form there is no B_0 since the means of deviation scores are always 0.

The partial derivatives of G with respect to B_1 and B_2 may be written as follows:

with $\frac{\delta G}{\delta B_1} = 0$ and $\frac{\delta G}{\delta B_2} = 0$ we obtain

$$B_1 \sum_{i=1}^{N} x^2_{i,1} + B_2 \sum_{i=1}^{N} x_{i,1}x_{i,2} = \sum_{i=1}^{N} y_i x_{i,1}$$

and

$$B_1 \sum_{i=1}^{N} x_{i,1}x_{i,2} + B_2 \sum_{i=1}^{N} x^2_{i,2} = \sum_{i=1}^{N} y_i x_{i,2}$$

or in matrix notation

$$[\, B_1 \;\; B_2 \,] \begin{vmatrix} \sum_{i=1}^{N} x^2_{i,1} & \sum_{i=1}^{N} x_{i,1}x_{i,2} \\ \sum_{i=1}^{N} x_{i,1}x_{i,2} & \sum_{i=1}^{N} x^2_{i,2} \end{vmatrix}$$

$$= [\, \sum_{i=1}^{N} y_i x_{i,1} \;\; \sum_{i=1}^{N} y_i x_{i,2} \,]$$

or simply

$$[\, B \,]' [\, x'x \,] = [y'x]'$$

and

$$[\, B \,]' = [y'x]' [x'x]^{-1} \tag{5.36}$$

Matrix Form for Normal Equations Using Standardardized Scores

The regression equation from above may be written in terms of standardized (z) scores as

$$z'_y = \beta_1 z_1 + \beta_2 z_2 \tag{5.37}$$

The function to be minimized is $G = \sum_{i=1}^{N} (z_y - z'_y)^2$.

We obtain the partial derivatives of G with respect to β_1 and β_2 as before and set them to zero. The equations obtained are then

$$\beta_1 \sum_{i=1}^{N} z^2_1 + \beta_2 \sum_{i=1}^{N} z_1 z_2 = \sum_{i=1}^{N} z_y z_1$$

and $$\beta_1 \sum_{i=1}^{N} z_1 z_2 + \beta_2 \sum_{i=1}^{N} z^2_2 = \sum_{i=1}^{N} z_y z_2$$

If we divide both sides of the above equations by N we obtain

$$\beta_1 + \beta_2 r_{1,2} = r_{y,1}$$

$$\beta_1 r_{1,2} + \beta_2 = r_{y,2}$$

or

$$[\,\beta_1\ \beta_2\,] \begin{vmatrix} 1 & r_{1,2} \\ r_{1,2} & 1 \end{vmatrix} = [\, r_{y,1}\ \ r_{y,2}\,]$$

or more simply as

$$[\beta\,]'\,[\,r_{xx}\,] = [\,r_{y,x}\,]'$$

and therefore

$$[\beta\,]' = [\,r_{y,x}\,]'\,[\,r_{x,x}\,]^{-1} \tag{5.38}$$

Equations in the previous discussion are general forms for solving the regression coefficients $B_1,...,B_{k+1}$ in raw score form, the $B_1,...,B_k$ coefficients in deviation score form or the $\beta_1,...,\beta_k$ coefficients in standardized score form. In each case, the B's or Betas are obtained by multiplication of an inverse matrix times the vector of cross-products or correlations. When there are more than two independent variables, the inverse of the matrix becomes laborious to obtain by hand. Computers are generally available however, which makes the chore of obtaining an inverse much easier.

You should remember that the independent variables must, in fact, be independent. That is, one independent variable cannot be a sum of one or more of the other independent variables. If the assumption of independence is violated, the inverse of the matrix may not exist! In some cases, although the variables are independent, they may nevertheless correlate quite highly among themselves. In such cases (high colinearity among independent variables), the computation of the inverse matrix may be difficult and result in considerable error. If the determinant of the matrix is very close to zero, your results should be held suspect!

We will see in latter sections that the inverse of the matrix of independent variable cross-products, deviation cross-products or correlations may be used to estimate the standard errors of regression coefficients and the covariances among the regression coefficients.

Hypothesis Testing in Multiple Regression

Testing the Significance of the Multiple Regression Coefficient

The multiple regression coefficient $R_{Y,12...k}$ is an index of the degree to which the dependent and weighted composite of the independent variables correlate. The square of the coefficient indicates the proportion of variance of the dependent variable which is predicted by the independent variables. The $R^2_{Y,12...k}$ may be obtained from

$R^2_{Y,1..k} = [\ \beta\ ]'\ [\ r_{y,x}\ ]$ that is

$$R^2 = \beta_1 r_{y,1} + \beta_2 r_{y,2} + ... + \beta_k r_{y,k} \tag{5.39}$$

Since R^2 is a sample statistic which estimates a population parameter, it may be expected to vary from sample to sample and has a standard error.

The total sum of squares of the dependent variable Y may be partitioned into two main sources of variability:

(1) The sum of squares due to regression with the independent variables (SS_{reg}) and

(2) The sum of squares due to error or unexplained variance (SS_e).

We may estimate these values by

(a) $SS_{reg} = SS_Y\ R^2_{Y.12...k}$ and
(b) $SS_e = SS_Y\ (1 - R^2_{Y.12...k})$

Associated with each of these sums of squares are degrees of freedom. For the SS_{reg} the degrees of freedom is the number of independent variables, K. For the SS_e the degrees of freedom are N - K - 1, that is, the degrees of freedom for the variance of Y minus the degrees of freedom for regression. Since the sum of squares for regression and error are independent, we may form an F-ratio statistic as

$$F = \frac{MS_{reg}}{MS_e} = \frac{Ss_{reg} / K}{Ss_e / (N\text{-}K\text{-}1)} \cdot \frac{R^2_{Y,1..k}}{(1 - R^2_{Y,1..k})} \frac{N\text{-}K\text{-}1}{K} \tag{5.40}$$

The probability of the F statistic for K and (N-K-1) degrees of freedom may be estimated or values for the tails obtained from tables of the F distribution. If the probability of obtaining an F statistic as large or larger than that calculated is less than the alpha level selected, the hypothesis that $R^2 = 0$ in the population may be rejected.

The Standard Error of Estimate

The following Fig. illustrates that for every combination of the independent variables, there is a distribution of Y scores. Since our prediction equation based on a sample of observations yields only a single Y value for each combination of the independent variables, there are obviously some predicted Y scores that are in error. We may estimate the variability of the Y scores at any combination of the X scores. The standard deviation of these scores for a given combination of X scores is called the Standard Error of Estimate. It is obtained as

$S_{Y.X} = (SS_e / (N-K-1))^{½}$ (5.41)

Testing the Regression Coefficients

Just as we may test the hypothesis that the overall multiple regression coefficient does not depart significantly from zero, so may we test the hypothesis that a regression coefficient B does not depart significantly from zero. Note that if we conclude that the coefficient does not depart from zero, we are concluding that the associated variable for that coefficient does not contribute significantly to explaining (predicting) the variance of Y.

The regression coefficients have been expressed both in raw score form (B's) and in standardized score form (β's). We may convert from one form to the other using

$B_j = \beta_j S_Y / S_j$ (5.42)

or $\beta_j = B_j S_j / S_Y$

Since these coefficients are sample statistics, they have a standard error. The standard error of a regression coefficient may be obtained as the square root of:

$$S^2_{B_j} = \frac{S^2_{Y.X}}{SS_{Xj} (1 - R^2_{j,1..(k\_1)}} \quad (5.43)$$

where $S^2_{Y.X}$ is the standard error of estimate and SS_X is the sum of squares for the jth variable,

$R^2_{j,1..(k-1)}$ is the squared multiple correlation of the jth independent variable regressed on the K-1 remaining independent variables.

In using the above method to obtain the standard errors of regression coefficients, it is necessary to obtain the multiple correlation of each independent variable with the remaining independent variables.

Another method of obtaining the standard errors of B's is through use of the inverse of the matrix of deviation score cross-products among the independent variables. We indicated this matrix as

$[x'x]^{-1}$

in our previous discussion. If we multiply this matrix by the variance of our error of estimate $S^2_{Y.X}$ the resulting matrix is the variance-covariance matrix of regression coefficients. That is

$[C] = S^2_{Y.X} [x'x]^{-1}$ (5.44)

The diagonal elements of [C], that is, $C_{1,1}$, $C_{2,2}$,...,$C_{k,k}$ are the variances of the B regression coefficients and the off-diagonal values are the covariances of the regression coefficients for independent variables.

To test whether or not the B_j regression coefficient departs significantly from zero, we may use either the t-test statistic or the F-test statistic. The t-test is

$$t = \frac{B_j}{\sqrt{C_{j,j}}} = \frac{B_j}{S_{Bj}} \text{ with N-K-1 degrees freedom.}$$

Since the t^2 is equivalent to the F test with one degree of freedom in the numerator, we can similarly use the F statistic with 1 and N-K-1 degrees of freedom where

$$F = \frac{B_j^2}{C_{j,j}} \tag{5.45}$$

A third method for examining the effect of a single independent variable is to ask whether or not the inclusion of the variable in the regression model contributes significantly to the increase in the SS_{reg} over the regression model in which the variable is absent. Since the proportion of variance of Y that is accounted for by regression is R^2, we can obtain the proportion of variance accounted for by a variable by

$$R^2_{Y,1..j..K} - R^2_{Y,1..(K\_1)} \tag{5.46}$$

The first R^2 equation (we will call it the FULL Model) contains all independent variables. The second (which we will call the restricted model) is the proportion of Y score variance predicted by all independent variables except the jth variable. The difference then is the proportion of variance attributable to the jth variable. The sum of squares of Y for the jth variable is therefore

$$SS_j = SS_Y(R^2_{Y,1..j..K} - R^2_{Y,1..(K\_1)}) \tag{5.47}$$

The mean square for this source of variability is the same as the SS since there is only 1 degree of freedom. The ratio of the MS_j to MS_e forms an F statistic with 1 and N-K-1 degrees of freedom. That is

$$F = \frac{MS_j}{MS_e} = \frac{SS_Y(R^2_{full}-R^2_{restricted}) / 1}{SS_Y(1 - R^2_{full}) / (N-K-1)} \tag{5.48}$$

$$= \frac{R^2_{full} - R^2_{restricted}}{1 - R^2_{full}} \cdot \frac{N-K-1}{1}$$

If the independent variable j does not contribute significantly (incrementally) to the variance of Y, the F statistic above will not be significant at the alpha decision level value.

Testing the Difference Between Regression Coefficients

Two variables may differ in the cost of collection. For example, an aptitude test may cost the student or institution more than obtaining a high school grade point average. In selecting one or the other independent variable to use in a regression model, there arises the question as to whether or not two regression coefficients differ significantly between themselves. Since the regression coefficients are sample statistics, the difference between two coefficients B_j and B_k is itself a sample statistic. The regression coefficients B are not independent of one another unless the independent variables themselves are uncorrelated. The standard error of the difference between two coefficients must therefore take into account not only the variance of each coefficient but also their covariance. The variance of differences between two regression coefficients may be obtained as

$$S^2_{B_j - B_k} = C_{j,j} + C_{k,k} - C_{j,k} \tag{5.49}$$

where $C_{j,j}$, $C_{k,k}$ and $C_{j,k}$ are elements of the [C] matrix.
The test for significance of difference between two regression coefficients is therefore

$$t_{(N-K-1)} = \frac{B_j - B_k}{\sqrt{[C_{j,j}+C_{k,k}-C_{j,k}]}} \qquad (5.50)$$

Stepwise Multiple Regression

A popular procedure for doing multiple regression by means of a computer program involves what is called the Stepwise Multiple Regression procedure. One independent variable at a time is added to the regression model. The independent variables are added in decreasing order of contribution to the variance of Y. Typically, these programs will select the variable X which has the highest simple correlation with Y. Next, each of the remaining variables is tested to see which contributes the most to an increase in the R^2 (or corresponding F statistic). That variable which most contributes is added next. This process is repeated until all variables are entered or none contribute to a significant increase at the alpha level selected. Unfortunately, a variable that has been previously entered may no longer contribute significantly after another variable is entered due to the co-variability among the independent variables. For this reason, additional tests may be made of variables already entered for deletion. Clearly, if the alpha level for entry is equal to the entry level for deletion, one may repeatedly add and delete variables ad-infinitum. For that reason, a different criterion for deletion (larger) is used than for entry. A better method examines all combinations of 1, 2, 3, ..., K variables for that combination which yields the maximum R2. Due to the large number of models computed, this method consumes very large quantities of computer time. It should also be noted that most analyses are performed on a sample of data selected from a population. As such the sample correlations and variable means and standard deviations may be expected to vary from sample to sample. The stepwise methods will "capitalize" on chance variations in the data. A replication of the analysis with another sample of data will typically yield a different order of entry into the model.

Cross and Double Cross Validation of Regression Models

Because sample data are used in obtaining the regression coefficients and in obtaining estimates of R2, the investigator may well wonder whether or not the estimates obtained are stable. If prediction is the purpose for obtaining the coefficients, the investigator is not likely to predict scores of Y for those subjects in the analysis - the actual values of Y are known for those subjects. More often the question revolves around the accuracy of prediction of Y for another sample of subjects. The scores for this sample are predicted on the basis of a previous sample. When the actual Y scores (e.g. GPA's) become available, the difference between the predicted and actual Y scores can then be obtained. The sum of squared differences is then calculated and the validation coefficient computed as

$$V = \frac{SS_Y - SS_e}{SS_Y}$$

This ratio of predicted sum of squares to total sum of squares is comparable to the R^2 obtained in the original sample. Usually, the value of V is considerably smaller than the R^2 obtained in the original analysis. If the number of cases available is large, the investigator may "split" his/her sample into two parts. A multiple regression analysis is completed with one part and the resulting regression model used to predict the Y scores for the other part. This cross-validation method provides an immediate indication of the accuracy to expect in use of the model. Another variation involves obtaining regression coefficients from each half of the sample and applying the respective models to the other half sample. Pooling the errors of prediction from both samples yields a doubly-crossed validation index. Unfortunately, there are many ways to split a sample of N subjects into two parts. Each "split" can yield a different estimate of R^2 and B coefficients. For this reason, several methods have been developed to estimate the "shrunken" R^2 which will taken into account the sampling variations. These methods utilize the degrees of freedom used in obtaining the R^2. One estimate commonly used is

$$\text{Adjusted } R^2 = R^2 - (1 - R^2)\frac{N - 1}{N - K - 1}$$

The relationship between R2 and the number of subjects (N) and predictors (K) can be readily understood. If the number of subjects equals K + 1, the R^2 will always be 1.0 (assuming some variance in the variables). The reason is that all subjects must fall on the regression line, plane or hyperplane and there is no "freedom" to vary about the plane. As the ratio of the number of subjects to the number of variables studied increases, this "over-fitting" of the data to the plane decreases. The larger the ratio of number of subjects to the number of variables, the closer will the regression statistics estimate the population values. Notice however, that R^2 is biased toward over-estimation. This bias becomes smaller and smaller as the ratio of subjects to variables increases.

Block Entry Multiple Regression

Several options exist in LazStats for completing the typical Ordinary Least Squares Multiple Regression (OLS MR.) One method lets the user enter predictors in "blocks" of one or more variables. The results for each block are obtained and the regression process ended when further inclusion of variables make no further contribution at the level of significance chosen. To illustrate, we will use a file labeled cansas.LAZ. You will see that we use this file for a number of different procedures. Basically, the file contains three physical measurements for individuals and three performance measurements. The last performance measure is labeled "jump" and is a measure of how far the individual can jump. This will be our dependent variable to be predicted by one or more of the other variables. The dialog for this procedure is shown below:

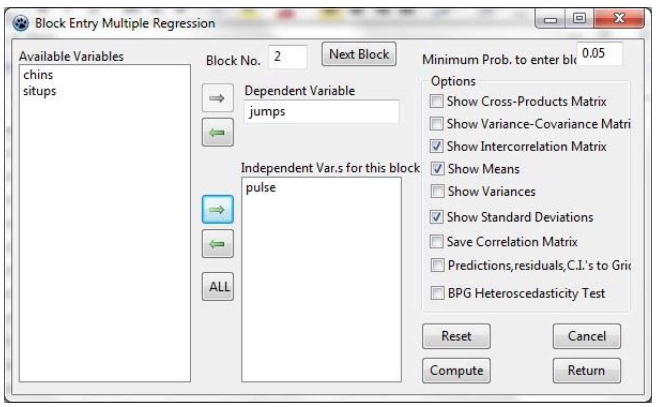


Fig. 5.4 Dialog for the Block Entry Multiple Regression Procedure

Two blocks have been entered in the above form. Two variables (weight and waist) were entered as the first block. Pulse is entered as the second block. When we click the Compute button we obtain:

```
Dependent variable: jumps

Variable       Beta      B        Std.Err.  t         Prob.>t   VIF       TOL
    weight    -0.246    -0.510    0.996    -0.513     0.615     4.121     0.243
     waist     0.022     0.359    7.679     0.047     0.963     4.121     0.243
 Intercept     0.000   148.772  146.699     1.014     0.325

SOURCE      DF        SS        MS        F       Prob.>F
Regression   2   2564.420  1282.210     0.460     0.6390
Residual    17  47393.780  2787.869
Total       19  49958.200

R2 = 0.0513, F =     0.46, D.F. = 2 17, Prob>F = 0.6390
Adjusted R2 = -0.0603

Standard Error of Estimate =    52.80
F =  0.460 with probability =  0.639
Block 1 did not meet entry requirements
```

Notice that neither of the variables in the first block contributed to a significant relationship to the dependent variable. The next block is then entered with the following results:

```
Dependent variable: jumps
```

```
Variable       Beta      B         Std.Err.  t         Prob.>t   VIF       TOL
    weight    -0.259    -0.538     1.034    -0.520     0.610     4.189     0.239
     waist     0.015     0.234     7.928     0.029     0.977     4.144     0.241
     pulse    -0.055    -0.389     1.863    -0.209     0.837     1.161     0.861
 Intercept     0.000   179.887   212.284     0.847     0.409

SOURCE      DF        SS        MS        F      Prob.>F
Regression  3    2692.894   897.631     0.304     0.8222
Residual    16   47265.306  2954.082
Total       19   49958.200

R2 = 0.0539, F =     0.30, D.F. = 3 16, Prob>F = 0.8222
Adjusted R2 = -0.1235

Standard Error of Estimate =    54.35
F =  0.304 with probability =  0.822
Block 2 did not meet entry requirements
```

Again, none of the variables predicted the jumps to a degree acceptable by most researchers (probabilities of the t and F values less than 0.05.)

Stepwise Forward Multiple Regression

Probably the most popular MR method taught and used by researchers is the stepwise procedure. In this procedure one variable at a time is added to the prediction model. In order to enter the equation, a variable must meet a user-specified significance level. In addition, to be retained in the model as other variables are entered, a previously entered variable must still be significant at another user-specified level (usually larger than the entry requirement.)

Shown below is the dialog used to predict jumps in the cansas.LAZ file using the three physical measurements of weight, waist and pulse.

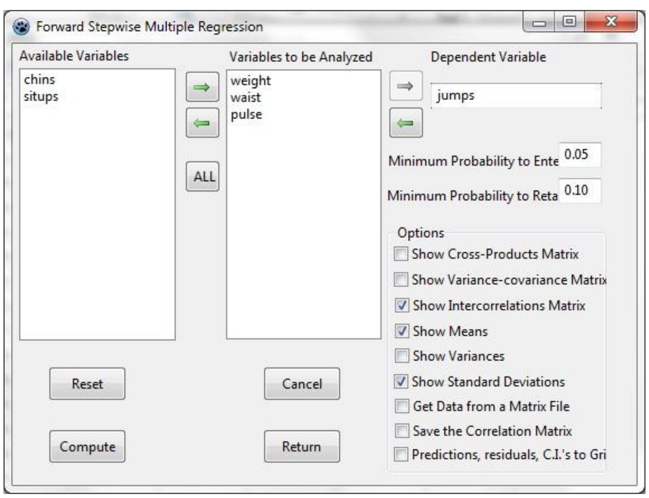


Fig. 5.5 Forward Stepwise MR Dialog

The obtained results are shown below:

```
Stepwise Multiple Regression by Bill Miller


Product-Moment Correlations Matrix

Variables     weight       waist       pulse       jumps
```

```
    weight      1.000       0.870      -0.366      -0.226
     waist      0.870       1.000      -0.353      -0.191
     pulse     -0.366      -0.353       1.000       0.035
     jumps     -0.226      -0.191       0.035       1.000


Means

Variables     weight       waist       pulse       jumps
              178.600      35.400      56.100      70.300




Standard Deviations

Variables     weight       waist       pulse       jumps
               24.691       3.202       7.210      51.277


Stepwise Multiple Regression by Bill Miller

----------------- STEP 1 ------------------
SOURCE    DF        SS      MS        F        Prob.>F
Regression   1  2558.343  2558.343     0.972     0.337
Residual    18 47399.857  2633.325
Total       19 49958.200

Dependent Variable: jumps

       R        R2         F     Prob.>F  DF1  DF2
   0.226     0.051     0.972     0.337    1   18
Adjusted R Squared = -0.002

Std. Error of Estimate =     51.316

Variable       Beta      B         Std.Error t         Prob.>t   VIF       TOL
    weight    -0.226    -0.470     0.477    -0.986     0.337     1.000     1.000
Exceeds limit - to be removed.

Constant =    154.237


Candidates for entry in next step.
Candidate  Partial  F Statistic  Prob.  DF1  DF2
weight      0.0000    0.0000    1.0000     1   18
waist       0.1238   -0.2716    1.0000     1   18
pulse       0.2295   -0.9009    1.0000     1   18
No further steps meet criterion for entry.

-------------FINAL STEP-----------
```

In this analysis, the first variable entered (weight) was not significant. The other two predictors made no significant additional contribution to the prediction of jumps and therefore the analysis was terminated. Notice that a test of the partial correlation was performed to determine whether or not an additional variable should be entered.

Backward Stepwise Multiple Regression

In contrast to the adding of one variable at a time as in the forward stepwise method, all predictors are initially entered and then eliminated, one by one in the backward method. This continues until all variables have been removed. The dialog for the analysis of the cansas.LAZ file is shown below:

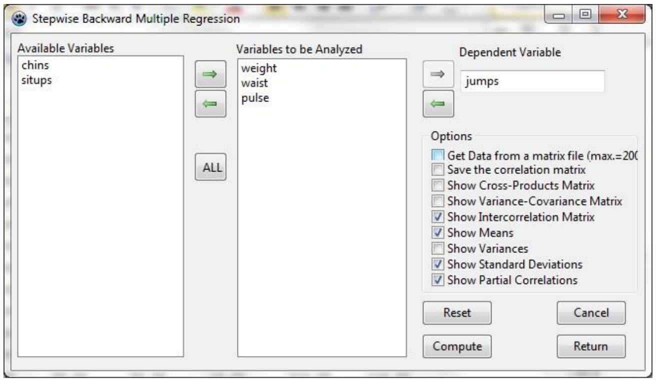


Fig. 5.6 Backward Stepwise MR Dialog

The results are shown below:

```
Backward Stepwise Multiple Regression by Bill Miller

----------------- STEP 1 ------------------
Determinant of correlation matrix =   0.1977

SOURCE     DF        SS       MS         F        Prob.>F
Regression   3  2692.894  897.631     0.304     0.822
Residual    16 47265.306  2954.082
Total       19 49958.200

Dependent Variable: jumps

       R        R2         F     Prob.>F  DF1  DF2
   0.232     0.054     0.304     0.822    3   16
Adjusted R Squared = -0.123

Std. Error of Estimate =     54.351

Variable       Beta      B         Std.Error t         Prob.>t   VIF       TOL
    weight    -0.259    -0.538     1.034     -0.520     0.610     4.189     0.239
     waist     0.015     0.234     7.928      0.029     0.977     4.144     0.241
     pulse    -0.055    -0.389     1.863     -0.209     0.837     1.161     0.861

Constant =    179.887

Partial Correlations

Variables      weight       waist       pulse
                -0.129       0.007      -0.052


Variable 2 (waist) eliminated


----------------- STEP 2 ------------------
Determinant of correlation matrix =   0.8196

SOURCE     DF        SS       MS         F        Prob.>F
Regression   2  2690.325  1345.162     0.484     0.625
Residual    17 47267.875  2780.463
Total       19 49958.200

Dependent Variable: jumps

       R        R2         F     Prob.>F  DF1  DF2
   0.232     0.054     0.484     0.625    2   17
Adjusted R Squared = -0.057

Std. Error of Estimate =     52.730

Variable       Beta      B         Std.Error t         Prob.>t   VIF       TOL
    weight    -0.246    -0.512     0.526     -0.972     0.344     1.154     0.866
```

```
     pulse    -0.055    -0.393     1.803    -0.218     0.830     1.154     0.866

Constant =    183.762

Partial Correlations

Variables     weight       pulse
               -0.230      -0.053


Variable 2 (pulse) eliminated


----------------- STEP 3 ------------------
Determinant of correlation matrix =   0.9488

SOURCE    DF        SS      MS        F        Prob.>F
Regression   1  2558.343  2558.343     0.972     0.337
Residual    18 47399.857  2633.325
Total       19 49958.200

Dependent Variable: jumps

       R        R2         F     Prob.>F  DF1  DF2
   0.226     0.051     0.972     0.337    1   18
Adjusted R Squared = -0.002

Std. Error of Estimate =     51.316

Variable       Beta      B         Std.Error t         Prob.>t   VIF       TOL
    weight    -0.226    -0.470     0.477    -0.986     0.337     1.000     1.000

Constant =    154.237

Partial Correlations

Variables     weight
               -0.226
```

You can see that in each step the variable with the smallest partial correlation is the one eliminated in the next step of the analysis.

Simultaneous Multiple Regression

Another method for obtaining multiple regression results is known as the simultaneous method. In this method the correlation among all of the included variables of the analysis are considered as the dependent variable to be predicted by the remaining entered variables. We will demonstrate by using all of the variables in the cansas.LAZ file as shown below:

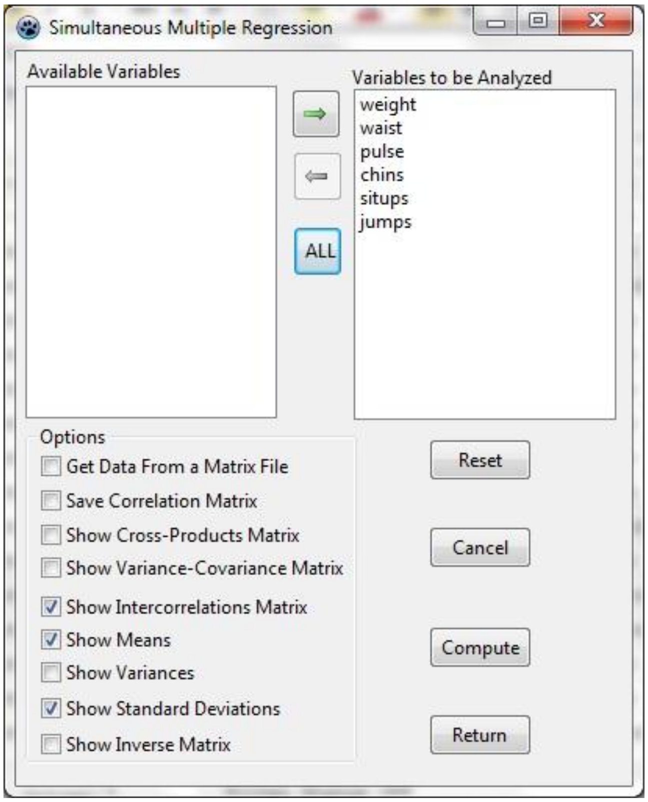


Fig. 5.7 Simultaneous MR Dialog

The results obtained are shown below. The first part gives the multiple correlation of each variable regressed on the other variables and the test of significance. Following that are the standardized regression coefficients (Betas) in columns for each dependent variable. The –1.000 in the column corresponds with the dependent variable. Also shown are the standard errors of prediction for each variable followed by the raw regression coefficients and intercepts (B weights) and the partial correlation of each variable with the remaining variables. It might be noted here that several factor analysis methods obtain simultaneous multiple regressions as a means of estimating how much common variance (the squared R) there is between each variable and the other variables.

```
Simultaneous Multiple Regression by Bill Miller

Product-Moment Correlations Matrix with   20 cases.


Variables
                 weight        waist        pulse        chins       situps
    weight        1.000        0.870       -0.366       -0.390       -0.493
     waist        0.870        1.000       -0.353       -0.552       -0.646
     pulse       -0.366       -0.353        1.000        0.151        0.225
     chins       -0.390       -0.552        0.151        1.000        0.696
    situps       -0.493       -0.646        0.225        0.696        1.000
     jumps       -0.226       -0.191        0.035        0.496        0.669


Variables
                  jumps
```

```
    weight      -0.226
     waist      -0.191
     pulse       0.035
     chins       0.496
    situps       0.669
     jumps       1.000



Means with   20 valid cases.

Variables       weight        waist        pulse        chins       situps
               178.600       35.400       56.100        9.450      145.550

Variables        jumps
                70.300


Standard Deviations with   20 valid cases.

Variables       weight        waist        pulse        chins       situps
                24.691        3.202        7.210        5.286       62.567

Variables        jumps
                51.277

Determinant of correlation matrix =   0.0208

Multiple Correlation Coefficients for Each Variable

  Variable       R        R2         F     Prob.>F  DF1  DF2
    weight     0.902     0.814    12.249     0.000    5   14
     waist     0.939     0.882    21.017     0.000    5   14
     pulse     0.386     0.149     0.490     0.778    5   14
     chins     0.734     0.539     3.275     0.036    5   14
    situps     0.884     0.782    10.026     0.000    5   14
     jumps     0.798     0.636     4.901     0.008    5   14

Betas in Columns with   20 cases.


Variables
                 weight        waist        pulse        chins       situps
    weight      -1.000        0.676       -0.321        0.347        0.372
     waist       1.070       -1.000        0.004       -0.616       -0.771
     pulse      -0.070        0.000       -1.000       -0.017        0.049
     chins       0.140       -0.157       -0.031       -1.000        0.143
    situps       0.317       -0.415        0.191        0.303       -1.000
     jumps      -0.301        0.317       -0.149        0.254        0.533


Variables
                  jumps
    weight      -0.588
     waist       0.982
     pulse      -0.064
     chins       0.201
    situps       0.888
     jumps      -1.000



Standard Errors of Prediction
Variable     Std.Error
    weight    12.407
     waist     1.279
     pulse     7.749
     chins     4.181
    situps    34.056
     jumps    36.020

Raw Regression Coefficients with   20 cases.
```

```
Variables
                 weight        waist        pulse        chins       situps
    weight      -1.000         0.088       -0.094         0.074         0.944
     waist       8.252        -1.000        0.008        -1.017       -15.069
     pulse      -0.240         0.000       -1.000        -0.012         0.424
     chins       0.655        -0.095       -0.042        -1.000         1.697
    situps       0.125        -0.021        0.022         0.026        -1.000
     jumps      -0.145         0.020       -0.021         0.026         0.650


Variables
                  jumps
    weight      -1.221
     waist      15.718
     pulse      -0.453
     chins       1.947
    situps       0.728
     jumps      -1.000



Variable   Constant
    weight  -114.302
     waist    22.326
     pulse    71.223
     chins    27.313
    situps   424.896
     jumps  -366.967

Partial Correlations with   20 cases.


Variables
                 weight        waist        pulse        chins       situps
    weight      -1.000         0.851       -0.150         0.221         0.344
     waist       0.851        -1.000        0.001        -0.311        -0.566
     pulse      -0.150         0.001       -1.000        -0.023         0.097
     chins       0.221        -0.311       -0.023        -1.000         0.208
    situps       0.344        -0.566        0.097         0.208        -1.000
     jumps      -0.420         0.558       -0.097         0.226         0.688


Variables
                  jumps
    weight      -0.420
     waist       0.558
     pulse      -0.097
     chins       0.226
    situps       0.688
     jumps      -1.000
```

Best Fit Multiple Regression

In many research projects an investigator is “searching” for possible relationships with a given variable of interest. As an exploratory method, this “best fit” method may be used to identify variables of possible interest for further exploration. The method finds the best combination of 2 or more predictor variables for explaining the variance of a dependent variable. Since one typically is using a sample, caution must be exercised because this method will “capitalize” on variations that normally occur from sample to sample.

To demonstrate this procedure, the file used in the other multiple regression procedures will be used (cansas.LAZ.) The dialog for this procedure is shown below:

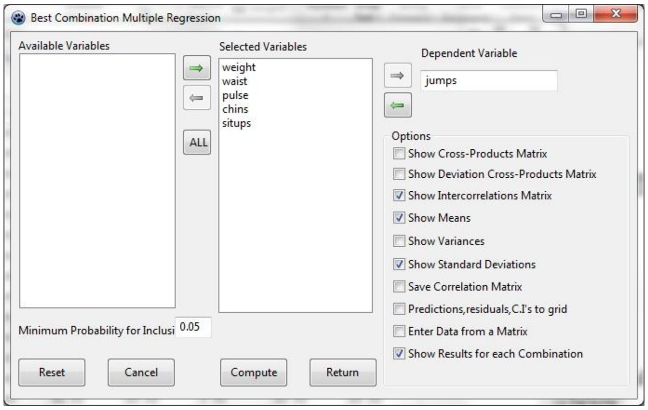


Fig. 5.8 Best Fit MR Dialog

The results obtained are:

```
Best Combination Multiple Regression by Bill Miller
 Set  1 includes variables:
variable 1 (weight)

Squared  R = 0.0512

 Set  1 includes variables:
variable 2 (waist)

Squared  R = 0.0367

 Set  1 includes variables:
variable 3 (pulse)

Squared  R = 0.0012

 Set  1 includes variables:
variable 4 (chins)

Squared  R = 0.2458

 Set  1 includes variables:
variable 5 (situps)

Squared  R = 0.4478


Variables entered in step  1
 5 situps

 Set  1 includes variables:
variable 5 (situps)

Squared  R = 0.4478

Squared Multiple Correlation = 0.4478
Dependent variable = jumps
ANOVA for Regression Effects :
SOURCE      df          SS           MS              F              Prob
Regression   1     22373.1193     22373.1193       14.5991          0.0013
Residual    18     27585.0807      1532.5045
Total       19     49958.2000

Variables in the equation
VARIABLE            b        s.e. b    Beta    t    prob. t
          situps    0.54846   0.1435   0.6692  3.821 0.0013
(Intercept)        -9.52819

Increase in squared R for this step = 0.447837
F =  14.5991 with D.F. 1 and 18 with Probability = 0.0013
----------------------------------------------------------
```

```
 Set  2 includes variables:
variable 1 (weight)
variable 2 (waist)

Squared  R = 0.0513

 Set  2 includes variables:
variable 1 (weight)
variable 3 (pulse)

Squared  R = 0.0539

 Set  2 includes variables:
variable 1 (weight)
variable 4 (chins)

Squared  R = 0.2471

 Set  2 includes variables:
variable 1 (weight)
variable 5 (situps)

Squared  R = 0.4620

 Set  2 includes variables:
variable 2 (waist)
variable 3 (pulse)

Squared  R = 0.0379

 Set  2 includes variables:
variable 2 (waist)
variable 4 (chins)

Squared  R = 0.2555

 Set  2 includes variables:
variable 2 (waist)
variable 5 (situps)

Squared  R = 0.5470

 Set  2 includes variables:
variable 3 (pulse)
variable 4 (chins)

Squared  R = 0.2474

 Set  2 includes variables:
variable 3 (pulse)
variable 5 (situps)

Squared  R = 0.4619

 Set  2 includes variables:
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.4496


Variables entered in step  2
 2 waist
 5 situps

 Set  2 includes variables:
variable 2 (waist)
variable 5 (situps)

Squared  R = 0.5470

Squared Multiple Correlation = 0.5470
Dependent variable = jumps
```

```
ANOVA for Regression Effects :
SOURCE      df          SS           MS            F             Prob
Regression   2     27329.4127     13664.7064       10.2657         0.0012
Residual    17     22628.7873      1331.1051
Total       19     49958.2000

Variables in the equation
VARIABLE           b        s.e. b    Beta    t    prob. t
           waist    6.60502   3.4230   0.4124  1.930 0.0705
          situps    0.76669   0.1752   0.9355  4.377 0.0004
(Intercept)      -275.10904

Increase in squared R for this step = 0.099209
F =   3.7234 with D.F. 1 and 17 with Probability = 0.0705
----------------------------------------------------------

 Set  3 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 3 (pulse)

Squared  R = 0.0539

 Set  3 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 4 (chins)

Squared  R = 0.3086

 Set  3 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 5 (situps)

Squared  R = 0.6125

 Set  3 includes variables:
variable 1 (weight)
variable 3 (pulse)
variable 4 (chins)

Squared  R = 0.2502

 Set  3 includes variables:
variable 1 (weight)
variable 3 (pulse)
variable 5 (situps)

Squared  R = 0.4696

 Set  3 includes variables:
variable 1 (weight)
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.4646

 Set  3 includes variables:
variable 2 (waist)
variable 3 (pulse)
variable 4 (chins)

Squared  R = 0.2556

 Set  3 includes variables:
variable 2 (waist)
variable 3 (pulse)
variable 5 (situps)

Squared  R = 0.5481

 Set  3 includes variables:
variable 2 (waist)
```

```
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.5577

 Set  3 includes variables:
variable 3 (pulse)
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.4636


Variables entered in step  3
 1 weight
 2 waist
 5 situps

 Set  3 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 5 (situps)

Squared  R = 0.6125

Squared Multiple Correlation = 0.6125
Dependent variable = jumps
ANOVA for Regression Effects :
SOURCE      df          SS          MS            F             Prob
Regression   3     30601.6275     10200.5425         8.4317         0.0014
Residual    16     19356.5725      1209.7858
Total       19     49958.2000

Variables in the equation
VARIABLE           b        s.e. b    Beta    t    prob. t
          weight   -1.09743   0.6673  -0.5284 -1.645 0.1195
           waist   14.61323   5.8617   0.9125  2.493 0.0240
          situps    0.81774   0.1699   0.9978  4.814 0.0002
(Intercept)      -370.02945

Increase in squared R for this step = 0.065499
F =   2.7048 with D.F. 1 and 16 with Probability = 0.1195
----------------------------------------------------------

 Set  4 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 3 (pulse)
variable 4 (chins)

Squared  R = 0.3098

 Set  4 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 3 (pulse)
variable 5 (situps)

Squared  R = 0.6168

 Set  4 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.6329

 Set  4 includes variables:
variable 1 (weight)
variable 3 (pulse)
variable 4 (chins)
variable 5 (situps)
```

```
Squared  R = 0.4719

 Set  4 includes variables:
variable 2 (waist)
variable 3 (pulse)
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.5583


Variables entered in step  4
 1 weight
 2 waist
 4 chins
 5 situps

 Set  4 includes variables:
variable 1 (weight)
variable 2 (waist)
variable 4 (chins)
variable 5 (situps)

Squared  R = 0.6329

Squared Multiple Correlation = 0.6329
Dependent variable = jumps
ANOVA for Regression Effects :
SOURCE      df          SS           MS           F             Prob
Regression   4     31619.5424      7904.8856        6.4658         0.0031
Residual    15     18338.6576      1222.5772
Total       19     49958.2000

Variables in the equation
VARIABLE           b        s.e. b    Beta    t    prob. t
          weight   -1.18963   0.6784  -0.5728 -1.754 0.0999
           waist   15.86516   6.0502   0.9907  2.622 0.0192
           chins    1.98471   2.1751   0.2046  0.912 0.3760
          situps    0.72449   0.1990   0.8840  3.641 0.0024
(Intercept)      -403.06414

Increase in squared R for this step = 0.020375
F =   0.8326 with D.F. 1 and 15 with Probability = 0.3760
----------------------------------------------------------


Last variable added failed entry test. Job ended.

Product-Moment Correlations Matrix with   20 cases.


Variables
                 weight        waist        pulse        chins       situps
    weight       1.000        0.870       -0.366       -0.390       -0.493
     waist       0.870        1.000       -0.353       -0.552       -0.646
     pulse      -0.366       -0.353        1.000        0.151        0.225
     chins      -0.390       -0.552        0.151        1.000        0.696
    situps      -0.493       -0.646        0.225        0.696        1.000
     jumps      -0.226       -0.191        0.035        0.496        0.669


Variables
                  jumps
    weight      -0.226
     waist      -0.191
     pulse       0.035
     chins       0.496
    situps       0.669
     jumps       1.000



Means with   20 valid cases.
```

```
Variables       weight        waist        pulse        chins       situps
               178.600       35.400       56.100        9.450      145.550

Variables        jumps
                70.300


Standard Deviations with   20 valid cases.

Variables       weight        waist        pulse        chins       situps
                24.691        3.202        7.210        5.286       62.567

Variables        jumps
                51.277
```

Polynomial (Non-Linear) Regression

In working with a variety of X and Y relationships, few investigators have failed to observe situations where the X and Y scores were not linearly related but rather were curvilinearly related. For example, achievement on a test may well have a curvilinear relationship with test anxiety - too little or too much producing a lower test score than a moderate degree of anxiety. To describe the relationship therefore requires the use of non-linear indices. We know from analytic geometry, that a curve may be described in cartesian coordinates by a polynomial in powers of X. For example, a parabola may be described by

$$Y = B_1X + B_2X^2 + B_0$$

In fact, a set of n data points (X,Y) can be completely "fit" by a polynomial of order n. Typically, however, we are interested in finding the lowest order (k) that adequately describes the Y variance. We could repeatedly obtain models with 1, 2, 3 .. n-1 terms each time obtaining the sum of squared residuals and stop adding values when the change in the error term was less than some arbitrary value. This could be done using the multiple regression programs already available in the multiple regression menu. Unfortunately, when values are raised to the power of 6 or higher, most computers suffer extensive "overflow" or round-off error in their calculations. To use higher order terms requires us to "transform" our data in such a manner that minimizes this problem. A popular method is to express each power of X in terms of an orthogonal polynomial $p_j(X)$

where

$$\sum_{i=1}^{n} p_j^2(X_i) = c_j \quad (j = 0,1..,q) \text{ and}$$

$$\sum_{i=1}^{n} p_j(X_i)p_k(X_i) = 0 \quad (j <\neq k)$$

for n variables of X

Solving these orthogonal polynomials and then transforming back to the original set of scores results in improving the degree of polynomial that can be analyzed.

Ridge Regression Analysis

Simple and multiple regression analyses using the least-squares method for estimating the regression coefficients assumes normally distributed, independent errors of the y scores corresponding to levels of the independent (X) scores. Frequently, these assumptions do not hold or there is high colinearity among the independent variables. In addition, the presence of "outliers" or extreme scores may often result in high distortion of the regression coefficients and their standard errors. There have been a variety of methods designed to provide alternative estimates of regression coefficients using criteria other than minimizing the squared differences of observed and predicted dependent scores. For example, one can attempt to minimize the absolute deviations or minimize the standard error of regression. One method which is finding increased use is termed "ridge regression".

In this method, the regression model (generalized ridge regression) is:

$Y = Z\beta' + e$
where Y is the vector of n dependent scores, Z is a matrix Z = XP where
X is the n by m matrix of independent scores
and P is the m by m matrix of eigenvectors of the X'X matrix, and
β' is the vector of coefficients estimated by
$\beta' = [Z'Z + K]^{-1}Z'Y$ where
K is a diagonal matrix of k_i values
with $k_i >= 0$, i = 1..m

The generalized ridge regression method minimizes the sum of squared deviations of the estimated coefficients β' from the values $\beta' = P'ß$ where ß is the vector of least-squares regression coefficients. The ridge regression analysis solves for an optimal set of k values. Even when the determinant of the X'X matrix nears zero (the rank of the matrix is less than the number of independent predictors), a set of coefficients will be obtained.

Binary Logistic Regression

(Contributed By John Pezzullo)

Background Info (just what is logistic regression, anyway?)

Ordinary regression deals with finding a function that relates a **continuous** outcome variable (dependent variable *y*) to one or more predictors (independent variables x_1, x_2, etc.). Simple linear regression assumes a function of the form:
$y = c_0 + c_1 * x_1 + c_2 * x_2 +...$
and finds the values of c_0, c_1, c_2, etc. (c_0 is called the "intercept" or "constant term").

Logistic regression is a variation of ordinary regression, useful when the observed outcome is **restricted to two values**, which usually represent the occurrence or non-occurrence of some outcome event, (usually coded as 1 or 0, respectively). It produces a formula that predicts the **probability of the occurrence** as a function of the independent variables.

Logistic regression fits a special s-shaped curve by taking the linear regression (above), which could produce any *y*-value between minus infinity and plus infinity, and transforming it with the function:
$p = \text{Exp}(y) / (1 + \text{Exp}(y))$ which produces *p*-values between 0 (as *y* approaches minus infinity) and 1 (as *y* approaches plus infinity). This now becomes a special kind of non-linear regression, which this page performs. Logistic regression also produces *Odds Ratios* (O.R.) associated with each predictor value. The *odds* of an event is defined as the probability of the outcome event **occurring** divided by the probability of the event **not occurring**. The odds ratio for a predictor tells the relative amount by which the odds of the outcome increase (O.R. greater than 1.0) or decrease (O.R. less than 1.0) when the value of the predictor value is increased by 1.0 units.

A standard iterative method is used to minimize the Log Likelihood Function (LLF), defined as the sum of the logarithms of the predicted probabilities of occurrence for those cases where the event occurred and the logarithms of the predicted probabilities of non-occurrence for those cases where the event did not occur.

Minimization is by Newton's method, with a very simple elimination algorithm to invert and solve the simultaneous equations. Central-limit estimates of parameter standard errors are obtained from the diagonal terms of the inverse matrix. Odds Ratios and their confidence limits are obtained by exponentiating the parameters and their lower and upper confidence limits (approximated by +/- 1.96 standard errors).

No special convergence-acceleration techniques are used. For improved precision, the independent variables are temporarily converted to "standard scores" (value - Mean) / StdDev. The *Null Model* is used as the starting guess for the iterations -- all parameter coefficients are zero, and the intercept is the logarithm of the ratio of the number of cases with *y*=1 to the number with *y*=0. Convergence is not guaranteed, but this page should work properly with most practical problems that arise in real-world situations.

This implementation has no predefined limits for the number of independent variables or cases. The actual limits are probably dependent on the user's available memory and other computer-specific restrictions.

When this analysis is selected from the menu, the form below is used to select the dependent and independent variables. We are using the BinaryReg2.LAZ file for an example.

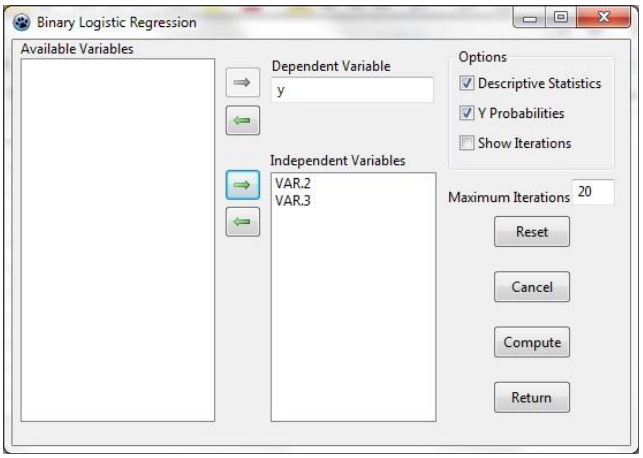


Fig. 5.9 Binary Logistic MR Dialog

The results obtained are:

```
Logistic Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/logistic.html

Descriptive Statistics
5 cases have Y=0; 6 cases have Y=1.
Variable  Label            Average     Std.Dev.
     1           VAR.2     3.8091     2.3796
     2           VAR.3     3.8727     1.9480
Descriptive Statistics
5 cases have Y=0; 6 cases have Y=1.
Variable  Label            Average     Std.Dev.
     1           VAR.2     3.8091     2.3796
     2           VAR.3     3.8727     1.9480

Converged
Descriptive Statistics
5 cases have Y=0; 6 cases have Y=1.
Variable  Label            Average     Std.Dev.
     1           VAR.2     3.8091     2.3796
     2           VAR.3     3.8727     1.9480

Converged

Overall Model Fit... Chi Square =  10.3070 with df =   2 and prob. =   0.0058

Coefficients and Standard Errors...
Variable        Label     Coeff.     StdErr     p
    1           VAR.2     1.2452     1.5087     0.4092
    2           VAR.3     2.2438     1.7746     0.2061
Overall Model Fit... Chi Square =  10.3070 with df =   2 and prob. =   0.0058

Coefficients and Standard Errors...
Variable        Label     Coeff.     StdErr     p
    1           VAR.2     1.2452     1.5087     0.4092
```

```
    2           VAR.3     2.2438     1.7746     0.2061
Intercept   -11.9555

Odds Ratios and 95% Confidence Intervals...
Variable            O.R.       Low   --   High
          VAR.2     3.4735     0.1805    66.8282
          VAR.3     9.4294     0.2910   305.5108

      X           X         Y        Prob
     1.1000      2.0000    0      0.0022
     3.2000      4.0000    0      0.7319
     1.5000      1.6000    0      0.0015
     2.9000      2.4000    0      0.0493
     3.3000      1.9000    0      0.0270
     4.2000      3.3000    1      0.6635
     2.4000      4.1000    1      0.5579
     1.5000      6.2000    1      0.9786
     6.0000      4.0000    1      0.9889
     6.8000      4.7000    1      0.9991
     9.0000      8.4000    1      1.0000

Classification Table
           Predicted
        ---------------
Observed    0      1     Total
        ---------------
   0    |   4  |   1  |   5  |
   1    |   0  |   6  |   6  |
        ---------------
Total   |   4  |   7  |  11
        ---------------
```

Cox Proportional Hazards Survival Regression

(Contributed by John Pezzullo)

This program analyzes survival-time data by the method of Proportional Hazards regression (Cox). Given survival times, final status (alive or dead) , and one or more covariates, it produces a baseline survival curve, covariate coefficient estimates with their standard errors, risk ratios, 95% confidence intervals, and significance levels.

A patient asked his surgeon what the odds were of him surviving an impending operation. The doctor replied they were 50/50 but he'd be all right because the first fifty had already died!!

Background Information (just what is Proportional Hazards Survival Regression, anyway?)

Survival analysis takes the survival times of a group of subjects (usually with some kind of medical condition) and generates a survival curve, which shows how many of the members remain alive over time. Survival time is usually defined as the length of the interval between diagnosis and death, although other "start" events (such as surgery instead of diagnosis), and other "end" events (such as recurrence instead of death) are sometimes used.

The major mathematical complication with survival analysis is that you usually do not have the luxury of waiting until the very last subject has died of old age; you normally have to analyze the data while some subjects are still alive. Also, some subjects may have moved away, and may be lost to follow-up. In both cases, the subjects were known to have survived for some amount of time (up until the time you last saw them), but you don't know how much longer they might ultimately have survived. Several methods have been developed for using this "at least this long" information to preparing unbiased survival curve estimates, the most common being the Life Table method and the method of Kaplan and Meier.

We often need to know whether survival is influenced by one or more factors, called "predictors" or "covariates", which may be categorical (such as the kind of treatment a patient received) or continuous (such as the patient's age, weight, or the dosage of a drug). For simple situations involving a single factor with just two values (such as drug vs placebo), there are methods for comparing the survival curves for the two groups of subjects. But for more complicated situations we need a special kind of regression that lets us assess the effect of each predictor on the shape of the survival curve.

To understand the method of proportional hazards, first consider a "baseline" survival curve. This can be thought of as the survival curve of a hypothetical "completely average" subject -- someone for whom each predictor variable is equal to the average value of that variable for the entire set of subjects in the study. This baseline survival curve doesn't have to have any particular formula representation; it can have any shape whatever, as long as it starts at 1.0 at time 0 and descends steadily with increasing survival time.

The baseline survival curve is then systematically "flexed" up or down by each of the predictor variables, while still keeping its general shape. The proportional hazards method computes a coefficient for each predictor variable that indicates the direction and degree of flexing that the predictor has on the survival curve. Zero means that a variable has no effect on the curve -- it is not a predictor at all; a positive variable indicates that larger values of the variable are associated with greater mortality. Knowing these coefficients, we could construct a "customized" survival curve for any particular combination of predictor values. More importantly, the method provides a measure of the sampling error associated with each predictor's coefficient. This lets us assess which variables' coefficients are significantly different from zero; that is: which variables are significantly related to survival.

The log-likelihood function is minimized by Newton's method, with a very simple elimination algorithm to invert and solve the simultaneous equations. Central-limit estimates of parameter standard errors are obtained from the diagonal terms of the inverse matrix. 95% confidence intervals around the parameter estimates are obtained by a normal approximation. Risk ratios (and their confidence limits) are computed as exponential functions of the parameters (and their confidence limits). The baseline survival function is generated for each time point at which an event (death) occurred.

No special convergence-acceleration techniques are used. For improved precision, the independent variables are temporarily converted to "standard scores" (value - Mean) / StdDev. The *Null Model* (all parameters = 0)is used as the starting guess for the iterations. Convergence is not guaranteed, but this page should work properly with most real-world data.

There are no predefined limits to the number of variables or cases this page can handle. The actual limits are probably dependent on your computer's available memory.

The specification form for this analysis is shown below with variables entered for a sample file labeled COXREGDATA.LAZ:

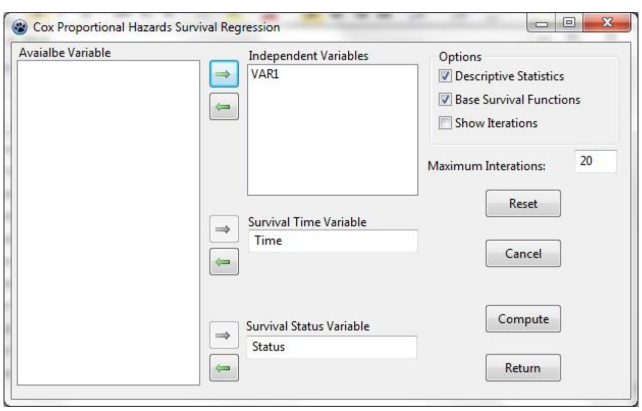


Fig. 5.10 Cox Proportional Hazzard Regression Dialog

Results for the above sample are as follows:

```
Cox Proportional Hazards Survival Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/prophaz.html
```

```
Descriptive Statistics
Variable  Label           Average     Std.Dev.
     1            VAR1    51.1818    10.9778

Converged

Overall Model Fit...
Chi Square =   7.3570 with d.f. 1 and probability =   0.0067

Coefficients, Std Errs, Signif, and Confidence Intervals

Var             Coeff.    StdErr       p      Lo95%    Hi95%
      VAR1     0.3770     0.2542   0.1379  -0.1211   0.8752

Risk Ratios and Confidence Intervals

Variable      Risk Ratio   Lo95%     Hi95%
      VAR1     1.4580     0.8859     2.3993

Baseline Survivor Function (at predictor means)...
    2.0000     0.9979
    7.0000     0.9820
    9.0000     0.9525
   10.0000     0.8310
```

Weighted Least-Squares Regression

For regressions with cross-section data (where the subscript "i" denotes a particular individual or firm at a point in time), it is usually safe to assume the errors are uncorrelated, but often their variances are not constant across individuals. This is known as the problem of heteroskedasticity (for "unequal scatter"); the usual assumption of constant error variance is referred to as homoskedasticity. Although the mean of the dependent variable might be a linear function of the regressors, the variance of the error terms might also depend on those same regressors, so that the observations might "fan out" in a scatter diagram.

Approaches to Dealing with Heteroskedasticity

. For known heteroskedasticity (e.g., grouped data with known group sizes), use weighted least squares (WLS) to obtain efficient unbiased estimates;

. Test for heteroskedasticity of a special form using a squared residual regression;

. Estimate the unknown heteroskedasticity parameters using this squared residual regression, then use the estimated variances in the WLS formula to get efficient
estimates of regression coefficients (known as feasible WLS); or

. Stick with the (inefficient) least squares estimators, but get estimates of standard errors which are correct under arbitrary heteroskedasticity.

In this procedure, the "residualization" method is used to obtain weights that will reduce the effect of heteroskedastic values. The method consists of four stages:

Step 1. Perform an Ordinary Least Squares (OLS) regression and obtain the residuals and squared residuals where the residual is the difference between the observed dependent variable and the predicted dependent variable value for each case.

Step 2. Regress the values of the squared residuals on the independent variables using OLS. The F test for the model is an indication of heteroskedasticity in the data.

Step 3. Obtain the reciprocal of the square root of the absolute squared residuals. These weights are then multiplied times all of the variables of the regression model.

Step 4. Obtain the OLS regression of the weighted dependent variable on the weighted independent variables. One can obtain the regression through the origin. If elected, each variable's values are converted to deviations from their respective mean before the OLS analysis is performed.

As an alternative, the user may use weights he or she has derived. These should be similar to the reciprocal values obtained in step 3 above. When these weights are used, they are multiplied times the values of each variable and step 4 above is completed.

Shown below is the dialog box for the Weighted Least Squares Analysis and an analysis of the cansas.OS4 data file.

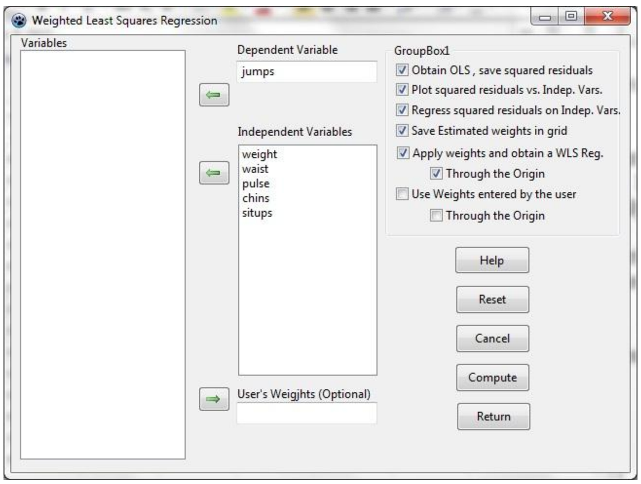


Fig. 5.11 Weighted Least Squares Regression Dialog

```
OLS REGRESSION RESULTS
Dependent variable: jumps


B WEIGHTS with   20 valid cases.

Variables       weight       waist       pulse       chins      situps
                -1.221      15.718      -0.453       1.947       0.728

Variables
              -366.967
MEANS with   20 valid cases.

Variables       weight       waist       pulse       chins      situps
               178.600      35.400      56.100       9.450     145.550

Variables
                70.300


VARIANCES with   20 valid cases.

Variables       weight       waist       pulse       chins      situps
               609.621      10.253      51.989      27.945    3914.576
```

```
Variables
              2629.379


STD. DEV.S with   20 valid cases.

Variables        weight        waist         pulse         chins        situps
                 24.691        3.202         7.210         5.286        62.567

Variables
                 51.277
Dependent variable: jumps


BETA WEIGHTS with   20 valid cases.

Variables        weight        waist         pulse         chins        situps
                 -0.588        0.982        -0.064         0.201         0.888
B STD.ERRORS with   20 valid cases.

Variables        weight        waist         pulse         chins        situps
                  0.704        6.246         1.236         2.243         0.205

Variables
                183.214


B t-test VALUES with   20 valid cases.

Variables        weight        waist         pulse         chins        situps
                 -1.734        2.517        -0.366         0.868         3.546

Variables
                 -2.003


B t VALUE PROBABILITIES with   20 valid cases.

Variables        weight        waist         pulse         chins        situps
                  0.105        0.025         0.720         0.400         0.003

Variables
                  0.065

SSY =   49958.20, SSreg =   31793.74, SSres =   18164.46
R2 = 0.6364, F =     4.90, D.F. = 5 14, Prob>F = 0.0084
Standard Error of Estimate =    36.02

REGRESSION OF SQUARED RESIDUALS ON INDEPENDENT VARIABLES
Dependent variable: ResidSqr


B WEIGHTS with   20 valid cases.

Variables        weight        waist         pulse         chins        situps
                -64.916      578.259       -50.564       124.826        16.375

Variables
```

```
             -8694.402
MEANS with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
               178.600        35.400        56.100         9.450       145.550

Variables
               908.196


VARIANCES with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
               609.621        10.253        51.989        27.945      3914.576

Variables
           4354851.627


STD. DEV.S with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
                24.691         3.202         7.210         5.286        62.567

Variables
              2086.828
Dependent variable: ResidSqr


BETA WEIGHTS with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
                -0.768         0.887        -0.175         0.316         0.491
B STD.ERRORS with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
                36.077       320.075        63.367       114.955        10.515

Variables
              9389.303


B t-test VALUES with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
                -1.799         1.807        -0.798         1.086         1.557

Variables
                -0.926


B t VALUE PROBABILITIES with   20 valid cases.

Variables       weight         waist         pulse         chins        situps
                 0.094         0.092         0.438         0.296         0.142

Variables
                 0.370
SSY = 82742180.90, SSreg = 35036253.36, SSres = 47705927.54
```

```
R2 = 0.4234, F =     2.06, D.F. = 5 14, Prob>F = 0.1323
Standard Error of Estimate =  1845.96
X versus Y Plot

X := ResidSqr, Y := weight from file:
C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable     Mean   Variance  Std.Dev.
ResidSqr    908.20  4354851.63   2086.83
weight      178.60    609.62     24.69
Correlation := -0.2973, Slope :=    -0.00, Intercept :=   181.79
Standard Error of Estimate :=    24.22
Number of good cases := 20
X versus Y Plot

X := ResidSqr, Y := waist from file:
C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable     Mean   Variance  Std.Dev.
ResidSqr    908.20  4354851.63   2086.83
waist        35.40     10.25      3.20
Correlation := -0.2111, Slope :=    -0.00, Intercept :=    35.69
Standard Error of Estimate :=     3.22
Number of good cases := 20
X versus Y Plot

X := ResidSqr, Y := pulse from file:
C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable     Mean   Variance  Std.Dev.
ResidSqr    908.20  4354851.63   2086.83
pulse        56.10     51.99      7.21
Correlation := -0.0488, Slope :=    -0.00, Intercept :=    56.25
Standard Error of Estimate :=     7.40
Number of good cases := 20
X versus Y Plot

X := ResidSqr, Y := chins from file:
C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable     Mean   Variance  Std.Dev.
ResidSqr    908.20  4354851.63   2086.83
chins         9.45     27.94      5.29
Correlation := 0.4408, Slope :=     0.00, Intercept :=     8.44
Standard Error of Estimate :=     4.88
Number of good cases := 20
X versus Y Plot

X := ResidSqr, Y := situps from file:
C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable     Mean   Variance  Std.Dev.
ResidSqr    908.20  4354851.63   2086.83
situps      145.55   3914.58     62.57
Correlation := 0.4775, Slope :=     0.01, Intercept :=   132.55
Standard Error of Estimate :=    56.48
Number of good cases := 20
```

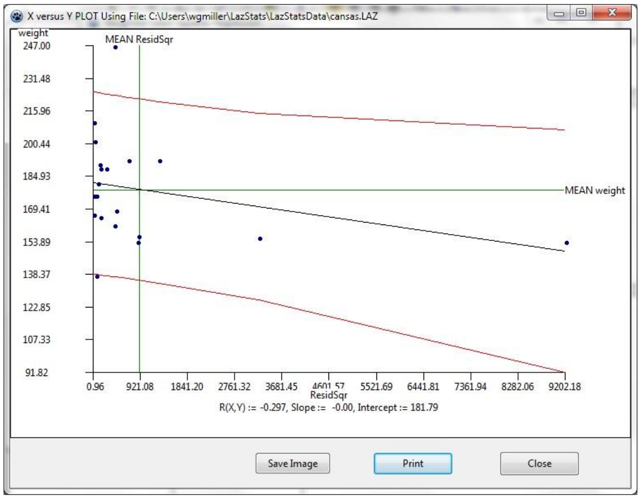


Fig. 5.12 Plot of Ordinary Least Squares Regression

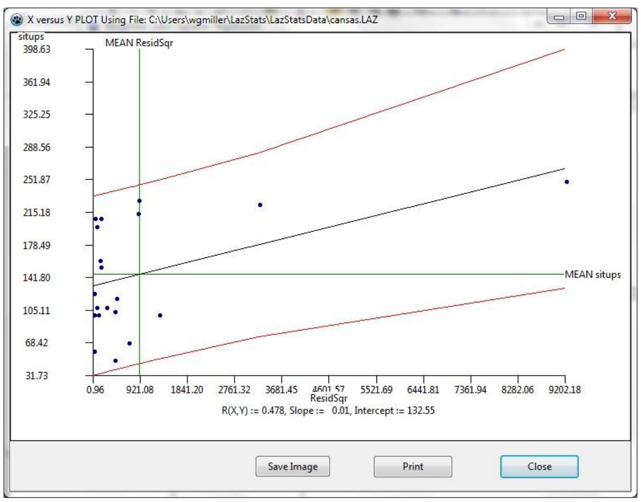


Fig. 5.13 Plot of Weighted Least Squares Regression

2-Stage Least-Squares Regression

Two Stage Least Squares regression may be used in the situation where the errors of independent and dependent variables are known (or likely) to be correlated. For example, the market price of a commodity and the demand for that commodity are non-recursive, that is, demand affects price and price affects demand. Prediction variables are "explanatory" variables to explain variability of the dependent variable. However, there may be other "instrumental" variables that predict one or more of these explanatory variables in which the errors are not correlated. If we first predict the explanatory variables with these instrumental variables and use the predicted values, we reduce the correlation of the errors with the dependent variable.

In this procedure, the user first selects the dependent variable of the study. Next, the explanatory variables (predictors) are entered. Finally, the instrumental variables AND the explanatory variables affected by these instrumental variables are entered into the instrumental variables list.

The two stages of this procedure are performed as follows:

Stage 1. The instrumental variables are identified as those in the instrumental list that are not in the explanatory list. The explanatory variables that are listed in both the explanatory and the instrumental lists are those for which predictions are to be obtained. These predicted scores are referred to as "proxy" scores. The predictions are obtained by regressing each explanatory variable listed in both lists with all of the remaining explanatory variables and instrumental variables. The predicted scores are obtained and stored in the data grid with a "P_" appended to the beginning of the original predictor name.

Stage 2. Once the predicted values are obtained, an OLS regression is performed with the dependent variable regressed on the proxy variables and the other explanatory variables not predicted in the previous stage.

In the following example, the cansas.LAZ file is analyzed. The dependent variable is the height of individual jumps. The explanatory (predictor) variables are pulse rate, no. of chinups and no. of situps the individual completes. These explanatory variables are thought to be related to the instrumental variables of weight and waist size. In the dialog box for the analysis, the option has been selected to show the regression for each of the explanatory variables that produces the predicted variables to be used in the final analysis. Results are shown below:

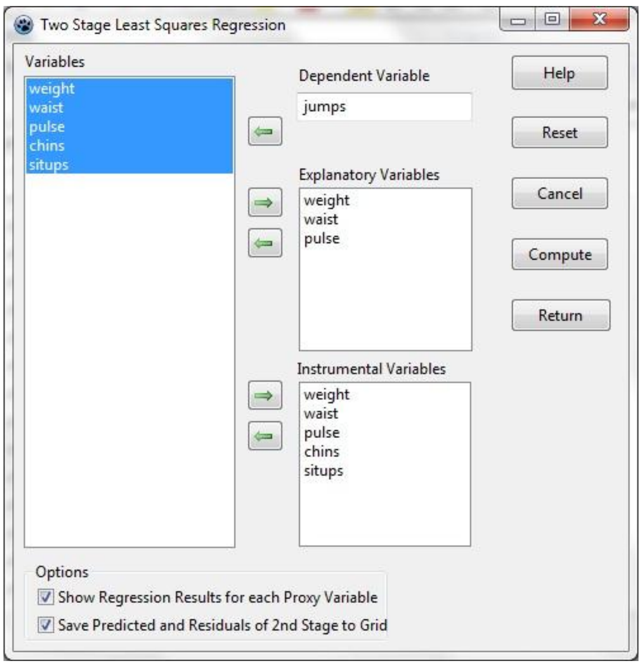


Fig. 5.14 Two Stage Least Squares Regression Dialog

```
FILE: C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Dependent := jumps
Explanatory Variables:
pulse
chins
situps
Instrumental Variables:
weight
waist
pulse
chins
situps
Proxy Variables:
P_pulse
P_chins
P_situps

Analysis for P_pulse
Dependent: pulse
Independent:
```

```
weight
waist
chins
situps
Dependent variable: pulse


B WEIGHTS with   20 valid cases.

Variables       weight        waist         chins       situps
                -0.069       -0.325        -0.084        0.007        79.673

MEANS with   20 valid cases.

Variables       weight        waist         chins       situps
               178.600       35.400         9.450      145.550        56.100


VARIANCES with   20 valid cases.

Variables       weight        waist         chins       situps
               609.621       10.253        27.945     3914.576        51.989


STD. DEV.S with   20 valid cases.

Variables       weight        waist         chins       situps
                24.691        3.202         5.286       62.567         7.210
Dependent variable: pulse


BETA WEIGHTS with   20 valid cases.

Variables       weight        waist         chins       situps
                -0.235       -0.144        -0.062        0.059

B STD.ERRORS with   20 valid cases.

Variables       weight        waist         chins       situps
                 0.146        1.301         0.468        0.043        32.257


B t-test VALUES with   20 valid cases.

Variables       weight        waist         chins       situps
                -0.471       -0.249        -0.179        0.158         2.470


B t VALUE PROBABILITIES with   20 valid cases.

Variables       weight        waist         chins       situps
                 0.644        0.806         0.860        0.876         0.026

SSY =     987.80, SSreg =     139.18, SSres =     848.62
R2 = 0.1409, F =     0.62, D.F. = 4 15, Prob>F = 0.6584
Standard Error of Estimate =     7.52
Analysis for P_chins
Dependent: chins
Independent:
weight
waist
pulse
situps
Dependent variable: chins


B WEIGHTS with   20 valid cases.

Variables       weight        waist         pulse       situps
                 0.045       -0.638        -0.026        0.047        18.641
MEANS with   20 valid cases.

Variables       weight        waist         pulse       situps
               178.600       35.400        56.100      145.550         9.450
```

```
VARIANCES with   20 valid cases.

Variables       weight        waist        pulse       situps
               609.621       10.253       51.989     3914.576       27.945


STD. DEV.S with   20 valid cases.

Variables       weight        waist        pulse       situps
                24.691        3.202        7.210       62.567        5.286

Dependent variable: chins


BETA WEIGHTS with   20 valid cases.

Variables       weight        waist        pulse       situps
                 0.208       -0.386       -0.035        0.557
B STD.ERRORS with   20 valid cases.

Variables       weight        waist        pulse       situps
                 0.080        0.700        0.142        0.020       20.533


B t-test VALUES with   20 valid cases.

Variables       weight        waist        pulse       situps
                 0.556       -0.911       -0.179        2.323        0.908


B t VALUE PROBABILITIES with   20 valid cases.

Variables       weight        waist        pulse       situps
                 0.586        0.377        0.860        0.035        0.378

SSY =     530.95, SSreg =     273.09, SSres =     257.86
R2 = 0.5143, F =     3.97, D.F. = 4 15, Prob>F = 0.0216
Standard Error of Estimate =     4.15

Analysis for P_situps
Dependent: situps
Independent:
weight
waist
pulse
chins
Dependent variable: situps


B WEIGHTS with   20 valid cases.

Variables       weight        waist        pulse        chins
                 0.284       -9.200        0.246        5.624      353.506

MEANS with   20 valid cases.

Variables       weight        waist        pulse        chins
               178.600       35.400       56.100        9.450      145.550


VARIANCES with   20 valid cases.

Variables       weight        waist        pulse        chins
               609.621       10.253       51.989       27.945     3914.576


STD. DEV.S with   20 valid cases.

Variables       weight        waist        pulse        chins
                24.691        3.202        7.210        5.286       62.567

Dependent variable: situps
```

```
BETA WEIGHTS with   20 valid cases.

Variables       weight        waist        pulse        chins
                 0.112       -0.471        0.028        0.475
B STD.ERRORS with   20 valid cases.

Variables       weight        waist        pulse        chins
                 0.883        7.492        1.555        2.421      211.726


B t-test VALUES with   20 valid cases.

Variables       weight        waist        pulse        chins
                 0.322       -1.228        0.158        2.323        1.670


B t VALUE PROBABILITIES with   20 valid cases.

Variables       weight        waist        pulse        chins
                 0.752        0.238        0.876        0.035        0.116
SSY =   74376.95, SSreg =   43556.05, SSres =   30820.90
R2 = 0.5856, F =     5.30, D.F. = 4 15, Prob>F = 0.0073
Standard Error of Estimate =    45.33
Dependent variable: jumps


B WEIGHTS with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                -3.794       11.381       -0.192      203.516
MEANS with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                56.100        9.450      145.550       70.300


VARIANCES with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                 7.325       14.373     2292.424     2629.379


STD. DEV.S with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                 2.706        3.791       47.879       51.277
Dependent variable: jumps


BETA WEIGHTS with   20 valid cases.

Variables      P_pulse      P_chins     P_situps
                -0.200        0.841       -0.179

B STD.ERRORS with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                 5.460        5.249        0.431      277.262


B t-test VALUES with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                -0.695        2.168       -0.445        0.734


B t VALUE PROBABILITIES with   20 valid cases.

Variables      P_pulse      P_chins     P_situps        chins
                 0.497        0.046        0.662        0.474
SSY =   49958.20, SSreg =   17431.81, SSres =   32526.39
R2 = 0.3489, F =     2.86, D.F. = 3 16, Prob>F = 0.0698
```

```
Standard Error of Estimate =    45.09
```

Non-Linear Regression as Implemented in LazStats

(Contributed From John Pezzullo's Non-Linear Regression page. http://members.aol.com/johnp71/nonlin.html)

Background Info (just what is nonlinear curve-fitting, anyway?):

Simple linear curve fitting deals with functions that are linear in the parameters, even though they may be nonlinear in the variables. For example, a parabola y=a+b*x+c*x*x is a nonlinear function of x (because of the x-squared term), but fitting a parabola to a set of data is a relatively simple linear curve-fitting problem because the parameters enter into the formula as simple multipliers of terms that are added together. Another example of a linear curve-fitting problem is y= a+b*Log(x)+c/x; the terms involve nonlinear functions of the independent variable x, but the parameters enter into the formula in a simple, linear way.

Unfortunately, many functions that arise in real world situations are nonlinear in the parameters, like the curve for exponential decay y=a*Exp(-b*x), where b is "wrapped up" inside the exponential function. Some nonlinear functions can be linearized by transforming the independent and/or dependent variables. The exponential decay curve, for example, can be linearized by taking logarithms: Log(y)=a'-b*x . The a' parameter in this new equation is the logarithm of a in the original equation,so once a' has been determined by a simple linear curve-fit, we can just take its antilog to get a.

But we often encounter functions that cannot be linearized by any such tricks, a simple example being exponential decay that levels off to some unknown value: y=a*Exp(-b*x)+c. Applying a logarithmic transformation in this case produces Log(y-c)=a'-b*x. This linearizes b, but now c appears inside the logarithm; either way, we're stuck with an intrinsically nonlinear parameter estimation problem, which is considerably more difficult than linear curve-fitting. That's the situation this web page was designed to handle.

For a more in-depth treatment of this topic, check out Dr. Harvey Motulsky's new web site: Curvefit.com -- a complete guide to nonlinear regression. Most of the information here is excerpted from Analyzing Data with GraphPad Prism, a book that accompanies the program GraphPad Prism. You can download this book as a pdf file.

Techie-stuff (for those who might be interested):

This procedure involves expanding the function to be fitted in a Taylor series around current estimates of the parameters, retaining first-order (linear) terms, and solving the resulting linear system for incremental changes to the parameters. The program computes finite-difference approximations to the required partial derivatives, then uses a simple elimination algorithm to invert and solve the simultaneous equations. Central-limit estimates of parameter standard errors are obtained from the diagonal terms of the inverse of the normal equations matrix. The covariance matrix is computed by multiplying each term of the inverse normal matrix by the weighted error-variance. It is used to estimate parameter error correlations and to compute confidence bands around the fitted curve. These show the uncertainty in the fitted curve arising from sampling errors in the estimated parameters, and do not include the effects of errors in the independent and dependent variables. The page also computes a generalized correlation coefficient, defined as the square root of the fraction of total y variance explainable by the fitted function.

Unequal weighting is accomplished by specifying the standard error associated with the y variable. Constant errors, proportional errors, or Poisson (square root) errors can be specified by a menu, and don't have to be entered with the data. Standard errors can also be entered along with the x and y variables. Finally, replicate y measurements can be entered; the program will compute the average and standard error of the mean.

Also available are a number of simple variable transformations (log, reciprocal, square root), which might simplify the function to be fitted and improve the convergence, stability and precision of the iterative algorithm. If a transformation is applied to the y variable, the program will adjust the weights appropriately.

The page also fits least-absolute-value curves by applying an iterative reweighting scheme by which each point is given a standard error equal to the distance of that point from the fitted curve. An option allows differential weighting of above-curve points vs. below-curve points to achieve a specified split of points above and below the curve (a percentile curve fit).

No special goal-seeking methods, precision-preserving techniques (such as pivoting), convergence-acceleration, or iteration-stabilizing techniques (other than a simple, user-specified fractional adjustment), are used. This method may not succeed with extremely ill-conditioned systems, but it should work with most practical problems that arise in real-world situations.

As an example, I have created a "parabola" function data set labeled parabola.TEX. To generate this file I used the equation y = a + b * x + c * x * x. I let a = 0, b = 5 and c = 2 for the parameters and used a sequence of x values for the independent variables in the data file that was generated. To test the non-linear fit program, I initiated the procedure and entered the values shown below:

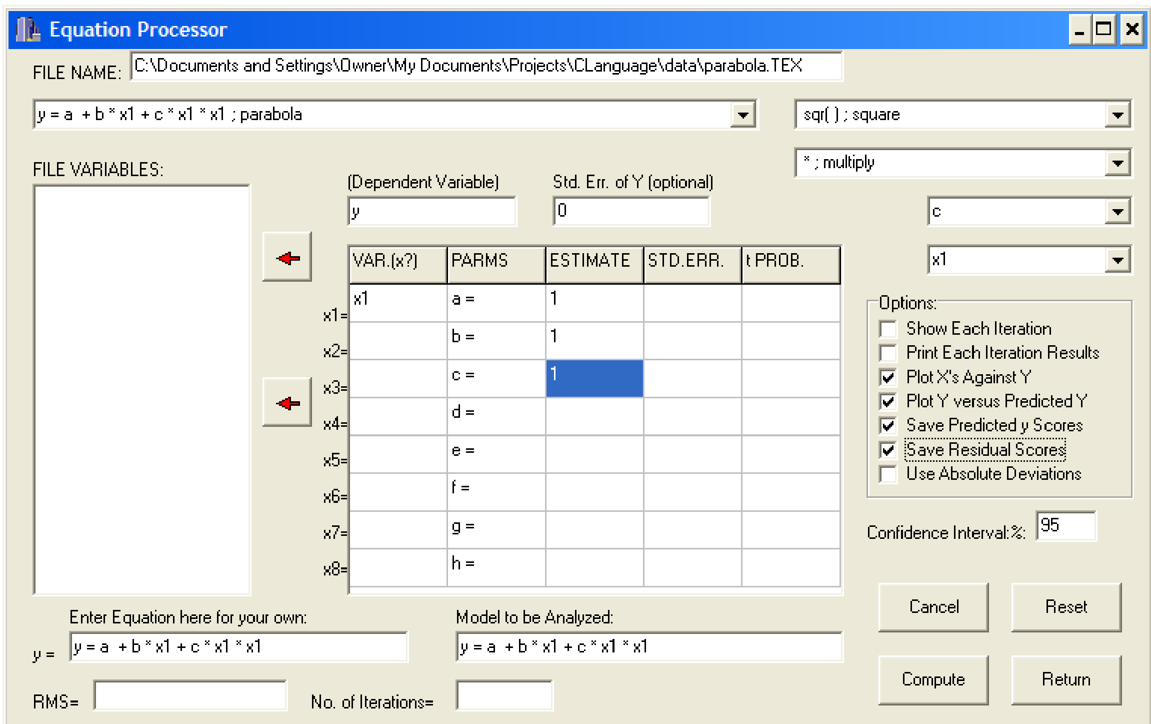


Fig. 5.15 Non Linear Regression Dialog

You can see that y is the dependent variable and x is the independent variable. Values of 1 have been entered for the initial estimates of a, b and c. The equation model was selected by clicking the parabola model from the drop-down models box. I could have entered the same equation by clicking on the equation box and typing the equation into that box or clicking parameters, math functions and variables from the drop-down boxes on the right side of the form. Notice that I selected to plot the x versus y values and also the predicted versus observed y values. I also chose to save the predicted scores and residuals (y - predicted y.)

The printed output shown below gives the model selected followed by the individual data points observed, their predicted scores, the residual, the standard error of estimate of the predicted score and the 95% confidence interval of the predicted score. These are followed by the obtained correlation coefficient and its square, root mean square of the y scores, the parameter estimates with their confidence limits and t probability for testing the significance of difference from zero.

```
y = a  + b * x1 + c * x1 * x1
     x4         y         yc        y-yc       SEest      YcLo       YcHi
    0.39800   2.31000   2.30863   0.00137   0.00161   2.30582   2.31143
   -1.19700  -3.13000  -3.12160  -0.00840   0.00251  -3.12597  -3.11723
   -0.48600  -1.95000  -1.95878   0.00878   0.00195  -1.96218  -1.95538
   -1.90800  -2.26000  -2.26113   0.00113   0.00522  -2.27020  -2.25205
   -0.84100  -2.79000  -2.79228   0.00228   0.00206  -2.79586  -2.78871
   -0.30100  -1.32000  -1.32450   0.00450   0.00192  -1.32784  -1.32115
```

```
    0.69600   4.44000   4.45208  -0.01208    0.00168    4.44917    4.45500
    1.11600   8.08000   8.07654   0.00346    0.00264    8.07195    8.08112
    0.47900   2.86000   2.85607   0.00393    0.00159    2.85330    2.85884
    1.09900   7.92000   7.91612   0.00388    0.00258    7.91164    7.92061
   -0.94400  -2.94000  -2.93971  -0.00029    0.00214   -2.94343   -2.93600
   -0.21800  -0.99000  -0.99541   0.00541    0.00190   -0.99872   -0.99211
    0.81000   5.37000   5.36605   0.00395    0.00183    5.36288    5.36923
   -0.06200  -0.31000  -0.30228  -0.00772    0.00185   -0.30549   -0.29907
    0.67200   4.26000   4.26629  -0.00629    0.00165    4.26342    4.26917
   -0.01900  -0.10000  -0.09410  -0.00590    0.00183   -0.09728   -0.09093
    0.00100   0.01000   0.00525   0.00475    0.00182    0.00209    0.00841
    0.01600   0.08000   0.08081  -0.00081    0.00181    0.07766    0.08396
    1.19900   8.88000   8.87635   0.00365    0.00295    8.87122    8.88148
    0.98000   6.82000   6.82561  -0.00561    0.00221    6.82177    6.82945

Corr. Coeff. =    1.00000  R2 =    1.00000
RMS Error =    5.99831, d.f. = 17  SSq =  611.65460

Parameter Estimates ...
p1=    0.00024  +/-     0.00182  p=    0.89626
p2=    5.00349  +/-     0.00171  p=    0.00000
p3=    2.00120  +/-     0.00170  p=    0.00000

Covariance Matrix Terms and Error-Correlations...
B(1,1)=             0.00000; r=   1.00000
B(1,2)=B(2,1)=     -0.00000; r=  -0.28318
B(1,3)=B(3,1)=     -0.00000; r=  -0.67166
B(2,2)=             0.00000; r=   1.00000
B(2,3)=B(3,2)=      0.00000; r=   0.32845
B(3,3)=             0.00000; r=   1.00000
X versus Y Plot

X = Y, Y = Y' from file: C:\Documents and Settings\Owner\My
Documents\Projects\CLanguage\data\parabola.LAZ

Variable     Mean   Variance  Std.Dev.
Y             1.76     16.29      4.04
Y'            1.76     16.29      4.04
Correlation = 1.0000, Slope =     1.00, Intercept =     0.00
Standard Error of Estimate =     0.01
Number of good cases = 20
```

Chapter 6. Analysis of Variance

Theory of Analysis of Variance

While the “Student” t-test provides a powerful method for comparing sample means for testing differences between population means, when more than two groups are to be compared, the probability of finding at least one comparison significant by chance sampling error becomes greater than the alpha level (rate of Type I error) set by the investigator. Another method, the method of Analysis of Variance, provides a means of testing differences among more than two groups yet retain the overall probability level of alpha selected by the researcher. Your LazStats4 package contains a variety of analysis of variance procedures to handle various research designs encountered in evaluation research. These various research designs require different assumptions by the researcher in order for the statistical tests to be justified. Fundamental to nearly all research designs is the assumption that random sampling errors produce normally distributed score distributions and that experimental effects result in changes to the mean, not the variance or shape of score distributions. A second common assumption to most designs using ANOVA is that the sub-populations sampled have equal score variances - this is the assumption of homogeneity of variance. A third common assumption is that the populations sampled have been randomly sampled and are very large (infinite) in size. A fourth assumption of some research designs where individual subjects or units of observation are repeatedly measured is that the correlation among these repeated measures is the same for populations sampled - this is called the assumption of homogeneity of covariance.

When we say we are "analyzing" variance we are essentially talking about explaining the variability of our values around the grand mean of all values. This "Total Sum of Squares" is just the numerator of our formula for variance. When the values have been grouped, for example into experimental and control groups, then each group also has a group mean. We can also calculate the variance of the scores within each of these groups. The variability of these group means around the grand mean of all values is one source of variability. The variability of the scores within the groups is another source of variability. The ratio of the variability of group means to the variability of within-group values is an indicator of how much our total variance is due to differences among our groups. Symbolically, we have "partioned" our total variability into two parts: variability among the groups and variability within the groups. We sometimes write this as

$$SS_T = SS_B + SS_W \qquad (6.1)$$

That is, the total sum of squares equals the sum of squares between groups plus the sum of squares within groups. The sums of squares are the numerators of variance estimates. The ratio of the SS_B to the SS_W forms our F test statistic. Later we will examine how we might also analyze the variability of scores using a linear equation.

The Completely Randomized Design

Introduction

Educational research often involves the hypothesis that means of scores obtained in two or more groups of subjects do not differ beyond that which might be expected due to random sampling variation. The scores obtained on the subjects are usually some measure representing relative amounts of some attribute on a dependent variable. The groups may represent different "treatment" levels to which subjects have been randomly assigned or they may represent random samples from some sub-populations of subjects that differ on some other attribute of interest. This treatment or attribute is usually denoted as the independent variable.

A Graphic Representation

To assist in understanding the research design that examines the effects of one independent variable (Factor A) on a dependent variable, the following representation is utilized:

```
___________________________________________________________________________
                           TREATMENT GROUP
         1              2              3              4              5
___________________________________________________________________________
```

Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
.	.	.	.	.
.	.	.	.	.
Y_{n1}	Y_{n2}	Y_{n3}	Y_{n4}	Y_{n5}

In the above Fig., each Y score represents the value of the dependent variable obtained for subjects 1, 2,...,n in groups 1, 2, 3, 4, and 5.

Null Hypothesis of the Design

When the researcher utilizes the above design in his or her study, the typical null hypothesis may be stated verbally as "the population means of all groups are equal". Symbolically, this is also written as

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k \qquad (6.2)$$

where k is the number of treatment levels or groups.

Summary of Data Analysis

The completely randomized design (or one-way analysis of variance design) depicted above requires the researcher to collect the dependent variable scores for each of the subjects in the k groups. These data are then typically analyzed by use of a computer program and summarized in a summary table similar to that below:

SOURCE	DF	SS	MS	F
Groups	k-1	$\sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2$	SS / k	$\frac{MS_g}{MS_e}$
Error	N-k	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2$	SS / (N-k)	
Total	N-1	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2$		

where Y_{ij} is the score for subject i in group j,

$\bar{Y}_j$ is the mean of scores in group j,

$\bar{Y}$ is the overall mean of scores for all subjects,

n_j is the number of subjects in group j, and

N is the total number of subjects across all groups.

Model and Assumptions

Use of the above research design assumes the following:

1. Variance of scores in the populations represented by groups 1,2,...,k are equal.

2. Error scores (which are the source of variability within the groups) are normally distributed.

3. Subjects are randomly assigned to treatment groups or randomly selected from sub-populations represented by the groups.

The model employed in the above design is

$$Y_{ij} = \mu + \mu_j + e_{ij} \qquad (6.3)$$

where μ is the population mean of all scores, μ_j is the effect of being in group j, and e_{ij} is the residual (error) for subject i in group j. In this model, it is assumed that the sum of the treatment effects (α_j) equals zero.

Fixed and Random Effects

In the previous section we introduced the analysis of variance for a single independent variable. In our discussion we indicated that treatment levels were usually established by the researcher. Those levels of treatment often are selected to represent specific intervals of a measurement on the independent variable, for example, amount of study time, level of drug dosage, time spent on a task, etc. The independent variable in many one-way analyses of variance may also represent classifications of objects or subjects such as political party, gender, grade level, or country of origin. We suggest more caution in interpretation of outcomes using classification variables since, in these cases, random assignment of subjects from a single population is usually impossible.

There is another situation for analysis of variance. That situation is where the researcher randomly selects levels of the independent variable (or works with objects that have random levels of an independent variable). For example, a researcher may wish to examine the effect of "amount of TV viewing" on student achievement. A random sample of students from a population might be drawn and those subjects tested. The subjects would also be asked to report the number of hours on the average that they watch TV during a week's time. If the analysis of variance is used, the variable "TV time" would be a random variable - the investigator has not assigned hour levels. If the experiment is repeated, the next sample of subjects would most likely represent different levels of TV time, thus the levels randomly fluctuate from sample to sample. For the one-way analysis of variance with the random effects model, the parameters estimated are the same as in the fixed effects model. For the one-way analysis of variance then, the analysis for the random-effects model is exactly the same as for the fixed-effects model (this will NOT be true for two-way and other higher level designs). An additional assumption of the random effects model is that the treatment effects (α) are normally distributed with mean 0 and variance σ_e^2. You may recognize that, if both dependent and independent variables are normally distributed and continuous, that the product-moment correlation may be an alternative method of analyzing data of the random-effects model.

Analysis of Variance - The Two-way, Fixed-Effects Design

A researcher may be interested in examining the effects of two (or more) independent variables on a given dependent variable at the same time. For example, a teacher may be interested in comparing the effects of three types of instruction, e.g. teacher lecture, small group discussion, and self instruction, on student achievement under two other conditions, e.g. students given a pretest and students not given a pretest. There is a possibility that both of these variables contribute to differences in student achievement. In addition, there is the possibility that method of instruction "interacts with" pre-testing conditions. For example, it might be suspected that use of a pretest with teacher lecture method is better than no pretest with teacher lecture but that such a difference would not be observed for the other two methods of instruction. The multi-factor ANOVA designs have the advantage of being able to examine not only the "main" effects of variables hypothesized to affect the dependent variable but also to be able to examine the interaction effects of those variables on the dependent variable.

The data may be conveniently depicted as a rectangle with the levels of one variable on the horizontal axis and the levels of the second variable on the vertical axis. The intersection of each row and column level is a treatment "cell" consisting of n_{jk} subjects receiving that combination of treatments. The table below gives the symbolic representation of scores in the two-way design:

		METHOD OF INSTRUCTION		
		Lecture	Group	Self
Pretest Condition	Pretest	$X_{111}=5$	$X_{112}=9$	$X_{113}= 5$
		$X_{211}=6$	$X_{212}=7$	$X_{213}=12$
		$X_{311}=4$	$X_{312}=6$	$X_{313}= 8$
	No Pretest	$X_{121}=10$	$X_{122}=6$	$X_{123}= 4$
		$X_{221}=12$	$X_{222}=8$	$X_{223}= 8$
		$X_{321}= 8$	$X_{322}=9$	$X_{323}= 5$

Using the above data it is possible to consider three seperate one-way ANOVA analyses:

1. An ANOVA of the three methods of instruction,
2. An ANOVA of the two pretesting conditions, and
3. An ANOVA of the 6 cells (treatment combinations).

The two-way ANOVA procedure yields all three in one analysis and provides greater sensitivity for each since the denominator of the F statistic will have the effects of the other two sources of variance already removed. The Summary table for the two-way ANOVA contains:

Source	D.F.	Sum of Squares	F	Parameters Estimated
Rows	R-1	$\sum_{j=1}^{R} n_{j.}(\bar{X}_{.j.}-\bar{X}_{...})^2$	MS_R/MS_e	$\sigma_e^2+\sigma_\alpha^2$
Columns	C-1	$\sum_{k=1}^{C} n_{.k}(\bar{X}_{..k}-\bar{X}_{...})^2$	MS_C/MS_β	$\sigma_e^2+\sigma_\beta^2$
Row x Col	(R-1)(C-1)	$\sum_{j=1}^{R}\sum_{k=1}^{C}(\bar{X}_{.jk}-\bar{X}_{.j.}-\bar{X}_{..k}+\bar{X}_{...})^2$		$\sigma_e^2+\sigma_{\alpha\beta}^2$

			MS_{RC}/MS_e
Error	$\sum_{j=1}^{R}\sum_{k=1}^{C}(n_{jk}-1)$	$\sum_{j=1}^{R}\sum_{k=1}^{C}\sum_{i=1}^{n_{jk}}(X_{ijk}-\bar{X}_{.jk})^2$	σ^2
Total	$N - 1$	$\sum_{j=1}^{R}\sum_{k=1}^{C}\sum_{i=1}^{n_{jk}}(X_{ijk}-\bar{X}_{...})^2$	

---

where X_{ijk} is the score for individual i in Row j and column k,

$\bar{X}_{.j.}$ is the mean of row j,

$\bar{X}_{..k}$ is the mean of column k,

$\bar{X}_{.jk}$ is the mean of the cell for row j and column k,

$\bar{X}_{...}$ is the overall (grand) mean.

As before, computational formulas may be developed from the defining formulas obtained from partitioning the total sum of squared deviations about the grand mean:

$$SS_T = \sum_{j=1}^{R}\sum_{k=1}^{C}\sum_{i=1}^{n_{jk}} X_{ijk}^2 - T_{...}^2/N \tag{6.4}$$

$$SS_R = \sum_{j=1}^{R} T_{.j.}^2/n_{j.} - T_{...}^2/N \tag{6.5}$$

$$SS_C = \sum_{k=1}^{C} T_{..k}^2/n_{.k} - T_{...}^2/N \tag{6.6}$$

$$SS_{RC} = \sum_{j=1}^{R}\sum_{k=1}^{C} T_{.jk}^2/n_{jk} - SS_R - SS_C - T_{...}^2/N \tag{6.7}$$

$$\sum_{j=1}^{R}\sum_{k=1}^{C}\sum_{i=1}^{n_{jk}} X_{ijk}^2 - \sum_{j=1}^{R}\sum_{k=1}^{C} T_{.jk}^2 / n_{jk} \tag{6.8}$$

where $T_{...}$ is the total of all scores,

$T_{.jk}$ is the total of the scores in a cell defined by the j row and k column,

$T_{.j.}$ is the total of the scores in the jth row,

$T_{..k}$ is the total of the scores in the kth column,

N is the total number of scores,

$n_{j.}$ is the number of scores in the jth row,

$n_{.k}$ is the number of scores in the kth column,

n_{jk} is the number of scores in the cell of the jth row and kth column.

In completing a two-way ANOVA, the researcher should attempt to have the same number of subjects in each group. If the ratio of any two columns is the same accross rows then the cell sizes are proportional and the analysis is still legitimate. If cell sizes are neither equal nor proportional, then the total sum of squares does not equal the sum of squares for rows, columns, interaction and error and the F tests do not represent independent tests of significance.

Stating the Hypotheses

The individual score of a subject (X_{ijk}) may be considered to be the linear composite of the effect of the row level (α_j), the effect of the column (β_k), the interaction effect of row and column combined ($\alpha\beta_{jk}$), the overall mean and random error, that is

$$X_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + e_{ijk} \tag{6.9}$$

The null hypotheses for the main effects therefore may be stated either as

$$H_o: \mu_{1.} = \mu_{2.} = ... = \mu_{j.} = ... \mu_{R.} \text{ for all rows,} \tag{6.10}$$

$$\text{or } H_o: \alpha_1 ... = \alpha_j = ... \alpha_R \text{ for all rows, and} \tag{6.11}$$

$$H_o: \mu_{.1} = ... = \mu_{.k} = ... = \mu_{.C} \text{ for all columns,} \tag{6.12}$$

$$\text{or } H_o: ß_1 = ... = ß_k = ... = ß_C \text{ for all columns, and} \tag{6.13}$$

$$H_o: (\mu_{jk}-\mu_{j.}-\mu_{.k}+\mu_{..}) \text{ for all row and column combinations,} \tag{6.14}$$

$$\text{or } H_o: \alpha ß_{11} = ... = \alpha ß_{jk} = ... = \alpha ß_{RC} \text{ for interactions.} \tag{6.15}$$

Again, we note that $\sum_{j=1}^{R} \alpha_{j.} = 0$, $\sum_{k=1}^{C} ß_{.k} = 0$ and $\sum_{j=1}^{R}\sum_{k=1}^{C} \alpha ß_{jk} = 0$. (6.16)

Interpreting Interactions

One may examine the means of cells in a two-way ANOVA using a plot such as illustrated in the Fig. below for our example of the teacher's research:

```
                        Plot of Cell Means
Achievement Score
12
11
10                 o
 9
 8                                  o                  x
 7                                  x
 6                                                     o
 5                 x
 4
 3
  ----------------------------------------------------------
            Lecture            Group       Self Instruction

      x = Pretest,          o = No Pretest
```

If lines are used to connect the o group means and lines are used to connect the x group means, one can see that the lines "cross".

If the lines for the pretest and no pretest levels are parallel across levels of the other factor, no interaction exists. When lines actually cross in the plot, this is called ordinal interaction. If the lines would cross if projected beyond current treatment levels, this is called disordinal interaction. In either case, the implication of interaction is that a particular combination of both treatments effects the dependent variable beyond the main effects alone. For example, if the interaction above is judged significant, then we cannot say that method 1 is better than method 3 of teaching without also specifying whether or not a pretest were used!

Note in the above interaction plot that the average of the three teaching method means are about the same for both pretest and no pretest conditions. This would indicate no main effect for the column variable pretest-no pretest. Similarly, the two means for each teaching method average about the same for each teaching method. This would indicate little effect of the variable teaching method (row). Your plot can graphically present effects due to the main variables as well as there interaction!

Random Effects Models

The two-way ANOVA design discussed to this point has assumed both factors contain fixed levels of treatment such that if the experiment was repeated numerous times, the levels would always be the same. If one or both of the factors represent random variables, that is, variables which would contain random levels upon replications of the experiment, then the expected values of the MS_{rows}, $MS_{columns}$, and $MS_{interaction}$ differ from that of the fixed-effects model. Presented below is a summary of the expected values for the two-way design when both variables are fixed, one variable random, and both variables random.

Both Row and Column Variables Fixed

Source	Expected MS	Calculated F-Ratio

Row	$\sigma^2_e + n_{j.}\sigma^2_\alpha$	MS_R / MS_e
Column	$\sigma^2_e + n_{.k}\sigma^2_\beta$	MS_C / MS_e
Interaction	$\sigma^2_e + n_{jk}\sigma^2_{\alpha\beta}$	MS_{RC} / MS_e
Error	σ^2_e	

Rows Fixed, Columns Random

Source	Expected MS	Calculated F-Ratio
Row	$\sigma^2_e + n_{..}\sigma^2_{\alpha\beta} + n_j\,\sigma^2_\alpha$	MS_R / MS_{RC}
Column	$\sigma^2_e + n_{.k}\,\sigma^2_\beta$	MS_C / MS_e
Interaction	$\sigma^2_e + n_{..}\sigma_{\alpha\beta}$	MS_{RC} / MS_e
Error	σ^2_e	

Row Random, Column Random

Source	Expected MS	Calculated F-Ratio
Row	$\sigma^2_e + n_{..}\sigma^2_{\alpha\beta} + n_{.j}\sigma^2_\alpha$	MS_R / MS_{RC}
Column	$\sigma^2_e + n_{..}\sigma^2_{\alpha\beta} + n_{k.}\sigma^2_\beta$	MS_C / MS_{RC}
Interaction	$\sigma^2_e + n_{..}\sigma^2_{\alpha\beta}$	MS_{RC} / MS_e
Error	σ^2_e	

One Between, One Repeated Design

Introduction

A common research design in education involves repeated measurements of several groups of subjects. For example, a pre- and post test administered to students in experimental and control courses may be considered a mixed design with one between subjects factor and one within subjects (repeated measures) factor. We might hypothesize that the means of the pretest equals the posttest, hypothesize that the experimental and control group means are equal and hypothesize that the change from pretest to post-test is the same for the two groups. Tests for these hypotheses use the F statistic.

As another example, suppose we are interested in the teacher evaluations given by three groups of administrators before and after three different teacher-evaluation training programs. All administrators are provided identical information on a sample of teachers including level and content of courses taught, school characteristics, community and student characteristics, and teacher characteristics such as degree, years experience, professional memberships, etc. plus a videotape of teaching excerpts. Each subject reviews all information and makes teacher ratings. The subjects are then randomly assigned to the three treatments: (1) a program on teacher evaluation which stresses the motivational aspects, (b) a program which stresses the teacher improvement aspect and (c) a program which stresses the reward aspect. Following these programs, each subject again evaluates the same or parallel teachers. The hypotheses tested would be that the mean teacher evaluations of each experimental group are equal, the mean evaluations prior to programs equal mean evaluations following the programs, and the change in mean teacher evaluations from pre to post program time are equal.

The Research Design

The Fig. below presents the schema for the mixed between and within factors design. Note that the different subjects in each "A" treatment group are repeatedly measured under each of the "B" treatment conditions. Our sample size is n subjects in each of M groups and the number of treatments is L.

The main hypotheses to be tested are

H_0: $\mu_{1.} = \mu_{2.} = ... = \mu_{M.}$ (all A levels are equal).
H_0: $\mu_{.1} = \mu_{.2} = ... = \mu_{.L}$ (all B levels are equal).
H_0: $\mu_{11} = \mu_{jk} = ... = \mu_{ML}$ (all AB cells are equal).

	B FACTOR TREATMENT LEVEL					
	1	2	3	4	L	Mean
A	X_{111}	X_{112}	X_{113}	X_{114}	X_{11L}	$\bar{X}_{11.}$
F	X_{211}	X_{212}	X_{213}	X_{214}	X_{21L}	$\bar{X}_{21.}$
A						
C	.	.	.	.	.	
T	.	.	.	.	.	$\bar{X}_{.1.}$
O						
R						
	X_{i11}	X_{i12}	X_{i13}	X_{i14}	X_{i1L}	$\bar{X}_{i1.}$
G						
R	.	.	.	.	.	.
O	.	.	.	.	.	.
U						
P	X_{n11}	X_{n12}	X_{n13}	X_{n14}	X_{n1L}	$\bar{X}_{n1.}$
1						
Mean	$\bar{X}_{.11}$	$\bar{X}_{.12}$	$\bar{X}_{.13}$	$\bar{X}_{.14}$	$\bar{X}_{.1L}$	$\bar{X}_{.1.}$
	X_{121}	X_{122}	X_{123}	X_{124}	X_{12L}	$\bar{X}_{12.}$
	X_{221}	X_{222}	X_{223}	X_{224}	X_{22L}	$\bar{X}_{22.}$
G	.	.	.	.	.	
R	.	.	.	.	.	$\bar{X}_{.2.}$
O						
U						
P	X_{i21}	X_{i22}	X_{i23}	X_{i24}	X_{i2L}	$\bar{X}_{i2.}$
2	.	.	.	.	.	.
	.	.	.	.	.	.
	X_{n21}	X_{n22}	X_{n23}	X_{n24}	X_{n2L}	$\bar{X}_{n2.}$
Mean	$\bar{X}_{.21}$	$\bar{X}_{.22}$	$\bar{X}_{.23}$	$\bar{X}_{.24}$	$\bar{X}_{.2L}$	$\bar{X}_{.2.}$
Col. Means	$\bar{X}_{..1}$	$\bar{X}_{..2}$	$\bar{X}_{..3}$	$\bar{X}_{..4}$	$\bar{X}_{..L}$	$\bar{X}_{...}$

Theoretical Model

The theoretical model for a subject i's score X from group j in Factor A on treatment k from factor B may be written

$$X_{ijk} = \mu + \alpha_j + ß_k + \pi_{i(j)} + ß_{jk} + ß\,\pi_{ki(j)} + e_{i(jk)} \qquad (6.17)$$

where μ is the population mean of the scores,
α_j is the effect of treatment j in Factor A,
$ß_k$ is the effect of treatment k in Factor B,
π_i is the effect of person i,
$\alpha ß_{jk}$ is the interaction of Factor A treatment j and treatment level k in Factor B,
$ß\pi_{ki(j)}$ is the interaction of subject i and B treatment k in the jth treatment group of A,
and $e_{i(jk)}$ is the error for person i in treatment j of Factor A and treatment k of Factor B.

In an experiment, we are usually interested in estimating the effect size of each treatment in each factor. We may also be interested in knowing whether or not there are significant differences among the subjects, and whether or not different subjects react differently to various treatments.

Assumptions

As in most ANOVA designs, we make a number of assumptions. For the mixed factors design these are:

1. The sum of treatment effects (α_j) is equal to zero,
2. The sum of treatment effects ($ß_k$) is equal to zero,
3. The sum of person effects ($\pi_{i(j)}$) is equal to zero,
4. The sum of $\alpha ß_{jk}$ interaction effects is equal to zero,
5. The sum of $ß\alpha_{ki(j)}$ interaction effects is equal to zero,
6. The sum of treatment x person interaction effects within levels of A ($ß\pi_{ki(j)}$) is zero,
7. The errors ($e_{i(jk)}$) are normally distributed with mean zero,
8. The variance of errors in each A treatment (α_j) are equal,
9. The variance of errors in each B treatment (β_k) are equal,
10. The covariances among the treatments ($COV_{pq(j)}$ p<>q p,q=1..L) within j levels of A are all equal.

The last assumption, equal covariances, means that if we were to transform scores within treatments to z scores, the correlations among the scores between any two treatments would all be equal in the population. You will also note that the denominator of the F ratios for testing differences among A treatment means is the pooled variance among subject means within groups as in a one-way ANOVA and the denominator of the F statistic for the Factors of B (the repeated measures) and the AxB interaction F statistic is the variance due to the pooled treatment by subjects interaction found in the Treatments by Subjects design.

Summary Table

The AxS ANOVA Summary table is often presented as follows:

SOURCE	D.F.	SS	MS	F
Between Subjects	Mn-1	$\sum_{j=1}^{M}\sum_{i=1}^{n} L(X_{ij.}-X_{...})^2$		

A	M-1	$nL\sum_{j=1}^{M}(\bar{X}_{.j.}-\bar{X}_{...})^2$	$SS_A/(M-1)$	MS_A/MS_{SwG}
Subjects within Groups	M(n-1)	$\sum_{j=1}^{M}\sum_{i=1}^{n} L(\bar{X}_{ij.}-\bar{X}_{.j.})^2$	$SS_{SwG}/[M(n-1)]$	
Within Subjects	Mn(L-1)	$\sum_{j=1}^{M}\sum_{k=1}^{L}\sum_{i=1}^{n}(X_{ijk}-\bar{X}_{.jk})^2$		
B	L-1	$nM\sum_{k=1}^{L}(\bar{X}_{..k}-\bar{X}_{...})^2$	$SS_B/(L-1)$	MS_B/MS_{BxSwG}
AxB	(M-1)(L-1)	$n\sum_{j=1}^{M}\sum_{k=1}^{L}(\bar{X}_{.jk}-\bar{X}_{..k}-\bar{X}_{.j.}+\bar{X}_{...})^2$	$SS_{AxS}/[(M-1)(L-1)]$	
BxS within Groups	M(n-1(L-1)	$\sum_{j=1}^{M} SS_{BS(j)}$	$SS_{BxSwG}/[M(n-1)(L-1)]$	
Total	nML-1	$\sum_{j=1}^{M}\sum_{k=1}^{L}\sum_{i=1}^{n}(X_{ijk}-\bar{X}_{..})^2$		

Population Parameters Estimated

The population mean of all scores (μ) is estimated by the overall mean. The mean squares provide estimates as follows:

MS_A estimates $\sigma_e^2 + M\sigma_\pi^2 + Mn\sigma_\alpha^2$
MS_{SwG} estimates $\sigma_e^2 + M\sigma_\pi^2$

MS_B estimates $\sigma_e^2 + \sigma_{\beta\pi}^2 + Mn\sigma^2_\beta$

MS_{AB} estimates $\sigma_e^2 + \sigma_{\beta\pi}^2 + n\sigma^2_{\alpha\beta}$

MS_{BxSwG} estimates $\sigma_e^2 + \sigma_{\beta\pi}^2$

Two Factor Repeated Measures Analysis

Repeated measures designs have the advantage that the error terms are typically smaller that designs using independent groups of observations. This was true for the Student t-test using matched or correlated scores. On the down-side, repeated measures on the same objects pose a special problem, particularly when the objects are human subjects. The main problem is "practice" or "learning" effects that may be greater for one treatment level than another. These effects are completely confounded with the actual treatment effects. While random or counter-balanced assignment of the treatments may reduce the cumulative effects to some degree, it does not remove the effects specific to a given treatment. It is also assumed that the covariance matrices are equal among the treatment levels. Users of these designs with human subjects should be careful to minimize the practice effects. This can sometimes be done by having subjects do tasks that are similar to those in the actual experiment before beginning trials of the experiment.

Nested Factors Analysis Of Variance Design

The Research Design

In the Nested ANOVA design, one factor (B) is completely nested within levels of another factor (A). Thus unlike the AxB Fixed Effects ANOVA in which all levels of B are crossed with all levels of A, each level of B is found in only one level of A in the nested design. The design may be graphically depicted as below:

A Factor	Treatment 1		Treatment j	Treatment M
B Factor	Level 1	Level 2	... Level k	.... Level L
Obser-vations	X_{111}	X_{112}	.... X_{1jk}	.... X_{1ML}
	X_{211}	X_{212}	.... X_{2jk}	.... X_{2ML}
	.	.	.... .	.... .
	.	.	.... .	.... .
	X_{n11}	X_{n12}	.... X_{njk}	.... X_{nML}
B Means	$\bar{X}_{.11}$	$\bar{X}_{.12}$	.... $\bar{X}_{.jk}$	.... $\bar{X}_{.ML}$
A Means	$\bar{X}_{.1.}$		$\bar{X}_{.j.}$	$\bar{X}_{.M.}$

The Variance Model

The observed X scores may be considered to be composed of several effects:

$$X_{ijk} = \mu + \alpha_j + \beta_{k(j)} + e_{k(j)} \tag{6.18}$$

The ANOVA Summary Table

We partition the total squared deviations of X scores from the grand mean of scores into sources of variation. The independent sources may be used to form F ratios for the hypothesis that the treatment means of A levels are equal and the hypothesis that the treatment levels of B are equal. The summary table (with sums of squares derivations) is as follows:

SOURCE	D.F.	SS	ESTIMATES:
A*	M-1	$\sum_{j=1}^{M} n_{j.}(\bar{X}_{.j.} - \bar{X}_{...})^2$	$\sigma^2_e + nD\sigma^2_B + nM\sigma^2_\alpha$
B (pooled)	$\sum_{j=1}^{M} (q_j - 1)$	$\sum_{j=1}^{M} \sum_{k=1}^{L_j} n_{jk}(\bar{X}_{.jk} - \bar{X}_{.j.})^2$	$\sigma^2_e + n\sigma^2_B$
Within	$\sum_{j=1}^{M} \sum_{k=1}^{L} (n_{jk} - 1)$	$\sum_{j=1}^{M} \sum_{k=1}^{L_j} \sum_{i=1}^{n_{jk}} (X_{ijk} - \bar{X}_{.jk})^2$	σ^2_e
Total N-1		$\sum_{j=1}^{M} \sum_{k=1}^{L_j} \sum_{i=1}^{n_{jk}} (X_{ijk} - \bar{X}_{...})^2$	

* Note: When factor B is a random effect, D = 1 and the F ratio for testing the A effect is the MS_A / MS_B . When factor B is a fixed effect, D=0 and the F ratio for testing A effects is MS_A / MS_w .

where:
X_{ijk} = An observed score in B treatment level k under A treatment level j ,

$\bar{X}_{.jk}$ = the mean of observations in B treatment level k in A treatment level j ,

$\bar{X}_{.j.}$ = the mean of observations in A treatment level j ,

$\bar{X}_{...}$ = the grand mean of all observations ,

n_{jk} = the number of observations in B treatment level k under A treatment level j

$n_{j.}$ = the number of observations in A treatment level j,

N = the total number of observations.

A, B and C Factors with B Nested in A

MODEL: $\bar{X}_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \varepsilon_{ijk}$ (6.19)

Assume that an experiment involves the use of two different teaching methods, one which involves instruction for 1 consecutive hour and another that involves two half-hours of instruction 4 hours apart during a given day. Three schools are randomly selected to provide method 1 and three schools are selected to provide method 2. Note that school is *nested* within method of instruction. Now assume that n subjects are randomly selected for each of two categories of students in each school. Category 1 students are males and category 2 students are female. This design may be illustrated in the table below:

	Instruction Method 1			Instruction Method 2		
	School 1	School 2	School 3	School 4	School 5	School 6
Category 1	n	n	n	n	n	n
Category 2	n	n	n	n	n	n

Notice that without School, the Categories are crossed with method and therefore are NOT nested. The expected values of the mean squares is:

Source of Variation	df	Expected Value
A (Method)	p - 1	$\sigma^2_e + nD_qD_r\sigma^2_{\beta\gamma} + nqD_r\sigma^2_{\alpha\gamma} + nrD_q\sigma^2_{\beta} + nqr\sigma^2_{\alpha}$
B within A	p(q-1)	$\sigma^2_e + nD_r\sigma^2_{\beta\gamma} + nr\sigma^2_{\beta}$
C (Category)	r - 1	$\sigma^2_e + nD_q\sigma^2_{\beta\gamma} + nqD_p\sigma^2_{\alpha\gamma} + npq\sigma^2_{\gamma}$
AC	(p-1)(r-1)	$\sigma^2_e + nD_q\sigma^2_{\beta\gamma} + nq\sigma^2_{\alpha\gamma}$
(B within A)C	p(q-1)(r-1)	$\sigma^2_e + n\sigma^2_{\beta\gamma}$
Within Cell	pqr(n-1) σ^2_e	

where there are p methods of A, q nested treatments B (Schools) and r C treatments (Categories). The D's with subscripts q, r or p have the value of 0 if the source is fixed and a value of 1 if the source is random. In this version of the analysis, all effects are considered fixed (D's are all zero) and therefore the F tests all use the Within Cell mean square as the denominator. If you use random treatment levels, you may need to calculate a more appropriate F test.

Latin and Greco-Latin Square Designs

Some Theory

In a typical 2 or 3-way analysis of variance design, there are independent groups assigned to each combination of the A, B (and C) treatment levels. For example, if one is designing an experiment with 3 levels of Factor A, 4 levels of Factor B and 2 levels of Factor C, then a total of 24 groups of randomly selected subjects would be used in the experiment (with random assignment of the groups to the treatment combinations.) With only 4 observations (subjects) per group, this would require 96 subjects in total. In such a design, one can obtain the main effects of A, B and C independent of the AxB, AxC, BxC and AxBxC interaction effects of the treatments. Often however, one may know before hand by previous research or by logical reasoning that the interactions should be minimal or would not exist. When such a situation exists, one can use a design which confounds or partially confounds such interactions with the main effects and drastically reduces the number of treatment groups required for the analysis. If the subjects can be repeatedly observed under various treatment conditions as in some of the previously discussed repeated-measures designs, then one can even further reduce the number of subjects required in the experiment. The designs to be discussed in this section utilize what are known as "Latin Squares".

The Latin Square

A Latin square is a balanced two-way classification scheme. In the following arrangement of letters, each letter occurs just once in each row and once in each column:

A	B	C
B	C	A
C	A	B

If we interchange the first and second row we obtain a similar arrangement with the same characteristics:

B	C	A
A	B	C
C	A	B

Two Latin squares are orthogonal if, when they are combined, the same pair of symbols occurs no more than once in the composite squares. For example, if the two Latin squares labeled Factor A and Factor B are combined to produce the composite shown below those squares the combination is NOT orthogonal because treatment combinations A1B2, A2B3, and A3B1 occur in more than one cell. However, if we combine Factor A and Factor C we obtain a combination that IS orthogonal.

FACTOR A			FACTOR B			FACTOR C		
A1	A2	A3	B2	B3	B1	C1	C2	C3
A2	A3	A1	B3	B1	B2	C3	C1	C2
A3	A1	A2	B1	B2	B3	C2	C3	C1

COMBINED A and B

A1B2	A2B3	A3B1
A2B3	A3B1	A1B2
A3B1	A1B2	A2B3

COMBINED A and C

A1C1	A2C2	A3C3
A2C3	A3C1	A1C2
A3C2	A1C3	A2C1

Notice that the 3 levels of treatment A and the 3 levels of treatment C are combined in such a way that no one combination is found in more than one cell. When two Latin squares are combined to form an orthogonal combination of the two treatment factors, the combination is referred to as a Greco-Latin square. Notice that the number of levels of both the treatment factors must be the same to form a square. Extensive tables of orthogonal Latin squares have been compiled by Cochran and Cox in "Experimental Designs", New York, Wiley, 1957.

Typically, the Greco-Latin square is represented using only the number (subscripts) combinations such as:

11	22	33
23	31	12
32	13	21

One can obtain additional squares by interchanging any two rows or columns of a Greco-Latin square. Not all Latin squares can be combined to form a Greco-Latin square. For example, there are no orthogonal squares for 6 by 6 or for 10 by 10 Latin squares. If the dimensions of a Latin square can be expressed as a prime number raised to the power of any integer n, then orthogonal squares exist. For example, orthogonal Latin squares exist of dimension 3, 4, 5, 8 and 9 from the relationships 3 from 3^1, 4 from 2^2, 5 from 5^1, 8 from 2^3, 9 from 3^2, etc.

Latin squares are often tabled in only "standard form". A square in standard form is one in which the letters of the first row and column are in sequence. For example, the following is a standard form for a 4 dimension square:

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

There are potentially a large number of standard forms for a Latin square of dimension n. There are 4 standard forms for a 4 by 4 square, and 9,408 standard forms for a 6 by 6 square. By interchanging rows and columns of the standard forms, one can create additional non-standard forms. For a 4 by 4 there are a total of 576 Latin squares and for a 6 by 6 there are a total of 812,851,200 squares! One can select at random a standard form for his or her design and then randomly select rows and columns to interchange to create a randomized combination of treatments.

Plan 1 by B.J. Winer

In his book "Statistical Principles in Experimental Design", New York, McGraw-Hill, 1962, Winer outlines a number of experimental designs that utilize Latin squares. He refers to these designs as "Plans" 1 through 13 (with some variations in several plans.) Not all plans have been included in OS2. Eight have been selected for inclusion at this time. The most simple design is that which provides the following model and estimates:

MODEL: $X_{ijkm} = \mu + \alpha_{i(s)} + \beta_{j(s)} + \gamma_{k(s)} + res_{(s)} + \varepsilon_{m(ijk)}$ (6.20)

Where i, j, k refer to levels of Factors A, B and C and m the individual subject in the unit. The (s) indicates this is a model from a Latin (s)quare design.

Source of Variation	Degrees of Freedom	Expected Mean Square
A	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\alpha$
B	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\beta$
C	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\gamma$
Residual	(p – 1)(p – 2)	$\sigma^2_\varepsilon + np\sigma^2_{res}$
Within cell	p2(n – 1)	σ^2_ε

In the above, p is the dimension of the square and n is the number of observations per unit.

Plan 2

Winer's Plan 2 expands the design of Plan 1 discussed above by adding levels of a Factor D. Separate Latin Squares are used at each level of Factor D. The plan of the design might appear as below:

		FACTOR B					FACTOR B		
	FACTOR	B1	B2	B3		FACTOR	B1	B2	B3
	A1	C3	C2	C1		A1	C1	C3	C2
FACTOR D1	A2	C1	C3	C2	FACTOR D2	A2	C2	C1	C3
	A3	C2	C1	C3		A3	C3	C2	C1

The analysis of Plan 2 is as follows:

Source of Variation	Degrees of Freedom	Expected Mean Square

A	p – 1	$\sigma^2_\varepsilon + npq\sigma^2_\alpha$
B	p – 1	$\sigma^2_\varepsilon + npq\sigma^2_\beta$
C	p – 1	$\sigma^2_\varepsilon + npq\sigma^2_\gamma$
D	q – 1	$\sigma^2_\varepsilon + npq\sigma^2_\delta$
AD	(p - 1)(q - 1)	$\sigma^2_\varepsilon + npq\sigma^2_{\alpha\delta}$
BD	(p - 1)(q - 1)	$\sigma^2_\varepsilon + npq\sigma^2_{\beta\delta}$
CD	(p - 1)(q - 1)	$\sigma^2_\varepsilon + npq\sigma^2_{\gamma\delta}$
Residual	q(p – 1)(p – 2)	$\sigma^2_\varepsilon + npq\sigma^2_{res}$
Within cell	$p^2q(n – 1)$	σ^2_ε

Notice that we can obtain the interactions with the D factor since all A, B and C treatments in the Latin square are observed under each level of D.   The model for Plan 2 expected value of the observed (X) score is:

$$Xijkmo = \mu + \alpha_{i(s)} + \beta_{j(s)} + \gamma_{k(s)} + \delta_m + \alpha\delta_{i(s)m} + \beta\delta_{j(s)m} + \gamma\delta_{k(s)m} + res_{(s)} \qquad (6.21)$$

As in Plan 1 described above, the (s) indicates sources from the Latin square.

Plan 3 Latin Squares Design

Plan 3 utilizes a balanced set of p x p Latin squares in a p x p x p factorial experiment. An example for a 3 x 3 x 3 design is shown below:

	FACTOR	FACTOR B B1	B2	B3
FACTOR D1	A1	C1	C2	C3
	A2	C2	C3	C1
	A3	C3	C1	C2

	FACTOR	FACTOR B B1	B2	B3
FACTOR D2	A1	C2	C3	C1
	A2	C3	C1	C2
	A3	C1	C2	C3

	FACTOR	FACTOR B B1	B2	B3
FACTOR D3	A1	C3	C1	C2
	A2	C1	C2	C3
	A3	C2	C3	C1

The levels of factors A, B and C are assigned at random to the symbols defining the Latin square. The levels of factor D are assigned at random to the whole squares. Notice the levels of each factor must be p, unlike the previous plan 2. In a complete 4 factor design wth three levels of each factor there would be 81 cells however with this design there are only 27. The main effect of factor D will be partially confounded with the ABC interaction however the main effects of A, B and C as well as the their interactions will be complete. The model of this design is:

$$E(X_{ijkmo}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \delta_m + \alpha\beta\gamma'_{ijk} \qquad (6.22)$$

The sources of variation, their degrees of freedom and parameter estimates are as shown below:

SOURCE	D.F.	E(MS)
A	$p - 1$	$\sigma^2_\varepsilon + np^2\sigma^2_\alpha$
B	$p - 1$	$\sigma^2_\varepsilon + np^2\sigma^2_\beta$
C	$p - 1$	$\sigma^2_\varepsilon + np^2\sigma^2_\gamma$
AB	$(p - 1)(p - 1)$	$\sigma^2_\varepsilon + np\sigma^2_{\alpha\beta}$
AC	$(p - 1)(p - 1)$	$\sigma^2_\varepsilon + np\sigma^2_{\alpha\gamma}$
BC	$(p - 1)(p - 1)$	$\sigma^2_\varepsilon + np\sigma^2_{\beta\gamma}$
D	$p - 1$	$\sigma^2_\varepsilon + np^2\sigma^2_\delta$
(ABC)’	$(p - 1)^3 - (p - 1)$	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma}$
Within cell	$p^3(n - 1)$	σ^2_ε

Analysis of Greco-Latin Squares

A Greco-Latin square design permits a three-way control of experimental units (row, column, and layer effects) through use of two Latin squares that are combined. One square is denoted with Latin letters and the other with Greek letters as illustrated below:

Square I			Square II			Combined Squares		
A	B	C	α	β	γ	Aα	Bβ	Cγ
B	C	A	γ	α	β	Bγ	Cα	Aβ
C	A	B	β	γ	α	Cβ	Aγ	Bα

Using numbers for the levels of the first and second effects, the composite square might also be represented by:

11	22	33
23	31	12
32	13	21

There are actually four variables: row, column, Latin-letter and Greek letter variables with p-squared cells in the composite square rather than p * p * p * p as there would be in a four-factor factorial design. The main effects of each of the factors will be confounded with the two-factor and higher interaction effects. Therefore this design is limited to the situations where the four factors are assumed to have negligible interactions. It is assumed that there are n independent observations in each cell.

The analysis that results provides the following sources of variation:

SOURCE	D.F.	E(MS)
A (Rows)	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\alpha$
B (Columns)	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\beta$
C (Latin Letters)	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\gamma$
D (Greek Letters)	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\delta$
Residual	(p – 1)(p – 3)	$\sigma^2_\varepsilon + n\sigma^2_{res}$
Within Cell	p2(n - 1)	σ^2_ε
Total	np2 – 1	

Plan 5 Latin Square Design

When the same unit (e.g. subject) may be observed under different treatment conditions, a considerable saving is realized in the sample size necessary for the experiment. As in all repeated measures designs however one must make certain assumptions about the homogeneity of variance and covariance. In plan 5 the levels of treatment under factor B are arranged in a Latin square with the columns representing levels of factor A. The rows are groups of subjects for which repeated measures are made across the columns of the square. The design is represented below:

```
               FACTOR A Levels
               A1     A2     A3
               -------------------------
         G1    B3     B1     B2
GROUP    G2    B1     B2     B3
         G3    B2     B3     B1
```

The model of the analysis is:

$$E(Xijkm) = \mu + \delta_k + \pi_{m(k)} + \alpha_i + \beta_j + \alpha\beta'_{ij} \quad (6.23)$$

The sources of variation are estimated by:

SOURCE	D.F.	E(MS)
Between Subjects	np – 1	
B	p – 1	$\sigma^2_\varepsilon + p\sigma^2_\pi + np\sigma^2_\delta$
Subjects in Groups	p(n-1)	$\sigma^2_\varepsilon + p\sigma^2_\pi$
Within Subjects	np(p-1)	
A	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\alpha$
B	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\beta$
(AB’)	(p – 1)(p – 2)	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta}$
error (within)	p(n – 1)(p – 1)	σ^2_ε

Plan 6 Latin Squares Design

Winer indicates that Plan 6 may be considered “as a fractional replication of a three-factor factorial experiment arranged in incomplete blocks.” Each subject within Group 1 is assigned to to treatement combinations abc_{111}, abc_{231} and abc_{321} such that each subject in the group is observed under all levels of factors A and B but under only one level of factor C. There is no balance with respect to any of the interactions but there is balance with respect to factors A and B. If all interactions are negligible relative to the main effects the following model and the sources of variation are appropriate:

$$E(X_{ijkm}) = \mu + \gamma_{k(s)} + \pi_{m(k)} + \alpha_{i(s)} + \beta_{j(s)} + res_{(s)}. \quad (6.24)$$

SOURCE OF VARIATION	D.F.	E(MS)
Between subjects	np – 1	
C	p – 1	$\sigma^2_\varepsilon + p\sigma^2_\pi + np\sigma^2_\gamma$
Subjects within groups	p(n – 1)	$\sigma^2_\varepsilon + p\sigma^2_\pi$
Within subjects	np(p – 1)	
A	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\alpha$
B	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\beta$
Residual	(p – 1)(p – 2)	$\sigma^2_\varepsilon + n\sigma^2_{res}$
Error (within)	p(n – 1)(p – 1)	σ^2_ε

The experiment may be viewed (for 3 levels of each variable) in the design below:

		LEVELS OF FACTOR A		
GROUP	LEVELS OF C	A1	A2	A3
G1	C1	B1	B3	B2
G2	C2	B2	B1	B3
G3	C3	B3	B2	B1

Plan 7 for Latin Squares

If, in the previous plan 6 we superimpose the Factors B and C as orthogonal Latin Squares, then Factor C is converted into a within-subjects effect. The Greco-Latin square design may be viewed as the following (for 3 levels of treatment):

	LEVELS OF FACTOR A		
Group	A1	A2	A3
G1	BC11	BC23	BC32
G2	BC22	BC31	BC13
G3	BC33	BC12	BC21

The expected value of X is given as:

$$E(X_{ijkmo}) = \mu + \delta_{m(s)} + \pi_{o(m)} + \alpha_{i(s)} + \beta_{j(s)} + \gamma_{k(s)} \quad (6.25)$$

The sources of variation, their degrees of freedom and the expected mean squares are:

SOURCE OF VARIATION	D.F.	E(MS)
Between subjects	np – 1	
Groups	p – 1	$\sigma^2_\varepsilon + p\sigma^2_\pi + np\sigma^2_\delta$
Subjects within groups	p(n – 1)	$\sigma^2_\varepsilon + p\sigma^2_\pi$
Within subjects	np(p – 1)	
A	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\alpha$
B	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\beta$
C	p – 1	$\sigma^2_\varepsilon + np\sigma^2_\gamma$
Residual	(p – 1)(p – 3)	$\sigma^2_\varepsilon + n\sigma^2_{res}$
Error (within)	p(n – 1)(p – 1)	σ^2_ε

Plan 9 Latin Squares

If we utilize the same Latin square for all levels of a Factor C we would have a design which looks like the outline shown below for 3 levels:

					LEVELS OF FACTOR C						
		C1				C2				C3	
	LEVELS OF FACTOR A				LEVELS OF FACTOR A				LEVELS OF FACTOR A		
GROUP	A1	A2	A3	GROUP	A1	A2	A3	GROUP	A1	A2	A3
G1	B2	B3	B1	G4	B2	B3	B1	G7	B2	B3	B1
G2	B1	B2	B3	G5	B1	B2	B3	G8	B1	B2	B3
G3	B3	B1	B2	G6	B3	B1	B2	G9	B3	B1	B2

The model for expected values of X is:

$$E(Xijkmo) = \mu + \gamma_k + (row)_m + (\gamma \text{ x } row)_{km} + \pi_{o(m)} + \alpha_i + \beta_j + \alpha\beta'_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma'_{ijk} \quad (6.26)$$

The sources of variation for Plan 9 are shown below:

SOURCE OF VARIATION	D.F.	E(MS)
Between subjects	npq - 1	
C	q – 1	$\sigma^2_\varepsilon + p\sigma_2 + np^2\sigma^2_\gamma$
Rows [AB(between)]	p – 1	$\sigma^2_\varepsilon + p\sigma_2 + nq\sigma^2_{\alpha\beta}$
C x row [ABC(between)]	(p – 1)(q – 1)	$\sigma^2_\varepsilon + p\sigma_2 + n\sigma^2_{\alpha\beta}$
Subjects within groups	pq(n – 1)	$\sigma^2_\varepsilon + p\sigma_2$

Within subjects	npq(p – 1)	
A	p – 1	$\sigma^2_\varepsilon + npq\sigma^2_\alpha$
B	p – 1	$\sigma^2_\varepsilon + npq\sigma^2_\beta$
AC	(p – 1)(q – 1)	$\sigma^2_\varepsilon + np\sigma^2_{\alpha\gamma}$
BC	(p – 1)(q – 1)	$\sigma^2_\varepsilon + np\sigma^2_{\beta\gamma}$
(AB)'	(p – 1)(p – 2)	$\sigma^2_\varepsilon + nq\sigma^2_{\alpha\beta}$
(ABC)'	(p – 1)(p – 3)(q – 1)	$\sigma^2_\varepsilon + n\sigma^2_{\alpha\beta\gamma}$
Error (within)	pq(p – 1)(n – 1)	σ^2_ε

In this design the groups and subjects within groups are considered random while, like previous designs, the A,B and C factors are fixed. Interactions with the group and subject effects are considered negligible.

One Fixed and One Random Factor ANOVA

Now let us run an example of an analysis with one fixed and one random factor. We will use the data file named “Threeway.LAZ which could also serve to demonstrate a three way analysis of variance (with fixed or random effects.) We will assume the row variable is fixed and the column variable is a random level. We select the One, Two and Three Way ANOVA option from the Comparisons sub-menu of the Statistics menu. The Fig. below shows how we specified the variables and their types:

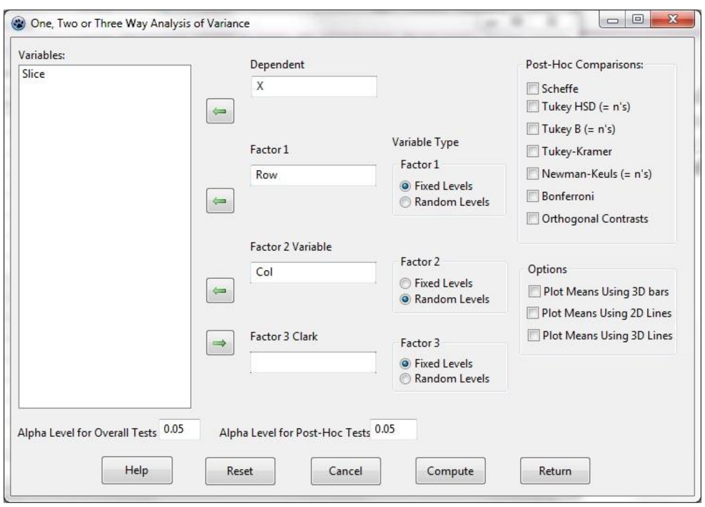


Fig. 6.1 Specification of a Two-Way ANOVA

Now when we click the Continue button we obtain:

```
Two Way Analysis of Variance

Variable analyzed: X

Factor A (rows) variable: Row (Fixed Levels)
Factor B (columns) variable: Col (Fixed Levels)

SOURCE          D.F.     SS         MS          F       PROB.> F   Omega Squared

Among Rows         1    12.250     12.250     5.765    0.022       0.074
Among Columns      1    42.250     42.250    19.882    0.000       0.293
Interaction        1    12.250     12.250     5.765    0.022       0.074
Within Groups     32    68.000      2.125
Total             35   134.750      3.850

Omega squared for combined effects =    0.441
```

```
Note: Denominator of F ratio is MSErr

Descriptive Statistics

GROUP Row Col.  N     MEAN   VARIANCE  STD.DEV.
Cell   1   1    9     3.000     1.500     1.225
Cell   1   2    9     4.000     1.500     1.225
Cell   2   1    9     3.000     3.000     1.732
Cell   2   2    9     6.333     2.500     1.581
Row    1       18     3.500     1.676     1.295
Row    2       18     4.667     5.529     2.351
Col    1       18     3.000     2.118     1.455
Col    2       18     5.167     3.324     1.823
TOTAL          36     4.083     3.850     1.962


TESTS FOR HOMOGENEITY OF VARIANCE
---------------------------------------------------------------------
Hartley Fmax test statistic =       2.00 with deg.s freedom: 4 and 8.
Cochran C statistic =       0.35 with deg.s freedom: 4 and 8.
Bartlett Chi-square statistic =       3.34 with    3 D.F. Prob. larger value
=  0.342
---------------------------------------------------------------------
```

You will note that the denominator of the F statistic for the two main effects may be different. You can also obtain plots for each main effect and the interaction effects.

Analysis of Variance - Treatments by Subjects Design

Introduction

A common research design in education involves repeated measurements of a group of subjects. For example, a test composed of K items administered to students in a course might be considered a "treatments by subjects" design. We might hypothesize that the means of the items are equal and test this hypothesis using the F statistic. As another example, suppose we are interested in changing teacher opinion about doing classroom research. We might develop a short attitude scale which measures their feelings concerning the feasibility and desirability of public school teachers conducting research. We may then design several "in-service" training programs and discussions concerned with classroom research. We administer our attitude instrument before the training programs, immediately following the training programs and a year later. The hypothesis tested is that the mean attitude at each of the three testing times is equal.

The Research Design

The Fig. below presents the schema for the Treatments by Subjects design. Note that the same subjects are measured under each of the "treatment" conditions. Our sample size is n subjects and the number of treatments is K.

The main hypothesis to be tested is H_0: $\mu_1 = \mu_2 = ... = \mu_k$.

```
______________________________________________________
         FACTOR TREATMENT GROUP
______________________________________________________
   1    2    3    4 .......... K         Mean
______________________________________________________
```

$$
\begin{array}{llllllll}
\text{S} & X_{11} & X_{12} & X_{13} & X_{14} & \ldots\ldots\ldots & X_{1k} & \bar{X}_{1.} \\
\text{U} & X_{21} & X_{22} & X_{23} & X_{24} & \ldots\ldots\ldots & X_{2k} & \bar{X}_{2.} \\
\text{B} & . & . & . & . & & . & . \\
 & . & . & . & . & & . & . \\
\text{J} & & & & & & & \\
 & X_{i1} & X_{i2} & X_{i3} & X_{i4} & \ldots\ldots\ldots & X_{ij} & \bar{X}_{i.} \\
\text{E} & & & & & & & \\
 & . & . & . & . & & . & . \\
\text{C} & . & . & . & . & & . & . \\
\text{T} & X_{n1} & X_{n2} & X_{n3} & X_{n4} & \ldots\ldots\ldots & X_{nk} & \bar{X}_{n.} \\
\hline
\text{Mean} & \bar{X}_{.1} & \bar{X}_{.2} & \bar{X}_{.3} & \bar{X}_{.4} & \ldots\ldots\ldots & \bar{X}_{.k} & \bar{X}_{..}
\end{array}
$$

Theoretical Model

The theoretical model for a subject i's score X on treatment j may be written

$$Xij = \mu + \alpha_j + \beta_i + \alpha\beta_{ij} + e_{ij}$$

where μ is the population mean of the scores,
α_j is the effect of treatment j,
β_I is the effect of person i,
$\alpha\beta_{ij}$ is the interaction of subject i and treatment j,
and e_{ij} is the error for person i in treatment j.

In an experiment, we are interested in estimating the effect size of each treatment. We may also be interested in knowing whether or not there are significant differences among the subjects, although this is usually not the case.

Summary Table

The Treatments by Subjects ANOVA Summary table is often presented as follows:

SOURCE	D.F.	SS	MS	F
A	K-1	$n\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{X}_{..})^2$	$SS_A/(K-1)$	MS_A/MS_{AxS}
Subjects	n-1	$K\sum_{i=1}^{n}(X_{i.}-\bar{X}_{..})^2$	$SS_S/(n-1)$	MS_S/MS_{AXS}
AxS Inter.	(K-1)(n-1)	$SST - SSA - SS_{AxS}$	$SS_{AxS}/[(K-1)(n-1)]$	
Total	Kn - 1	$\sum_{j=1}^{K}\sum_{i=1}^{n}(X_{ij}-\bar{X}_{..})^2$		

Assumptions

As in most ANOVA designs, we make a number of assumptions. For the Treatments by Subjects design these are:

1. The sum of treatment effects (α_j) are equal to zero,
2. The sum of person effects (β_i) are equal to zero,
3. The sum of treatment x person interaction effects ($\alpha\beta_{ij}$) are zero,
4. The errors (e_{ij}) are normally distributed with mean zero,
5. The variance of errors in each treatment (σ^2_j) are equal, and
6. The covariances among the treatments (COV_{jk}; $j<k$) are all equal.

The last assumption, equal covariances, means that if we were to transform scores within treatments to z scores, the correlations among the scores between any two treatments would all be equal in the population. You will also not that the denominator of the F ratios for testing differences among treatment means and among subject means is the treatment by subjects interaction rather than the usual within cell (pooled across cells) variance.

Population Parameters Estimated

The population mean of all scores (μ) is estimated by the overall mean. The mean squares provide estimates as follows:

MS_A estimates $\sigma^2_e + N\sigma^2_\alpha + \sigma^2_{\alpha\beta}$

MS_S estimates $\sigma^2_e + K\sigma^2_\beta$

MS_{AxS} estimates $\sigma^2_e + \sigma^2_\beta$

Computational Formulas

The algebraic formulas presented in the ANOVA Summary table above are not usually the most convenient for calculation of the sums of squares terms. The following formulas are usually used:

$$SS_A = \sum_{j=1}^{K} T_{j.}^2 / n - T_{..}^2/N$$

$$SS_S = \sum_{i=1}^{n} T_{i.}^2 / K - T_{..}^2 / N$$

$$SS_T = \sum_{j=1}^{K} \sum_{i=1}^{n} X_{ij}^2 - T_{..}^2 / N$$

$$SS_{AxS} = SS_T - SS_A - SS_S$$

where $T_j.$ is the total of score values within treatment j,
$T_i.$ is the total of score values for subject i,
T.. is the grand total of all score values,
n is the number of subjects, and
N is the grand number of all scores and equal to Kn.

An Example

To perform a Treatments by Subjects analysis of variance, we will use a sample data file labeled "ABRData.LAZ". We open the file and select the option "Within Subjects Anova" in the Comparisons sub-menu under the Statistics menu. The Fig. below is then completed as shown:

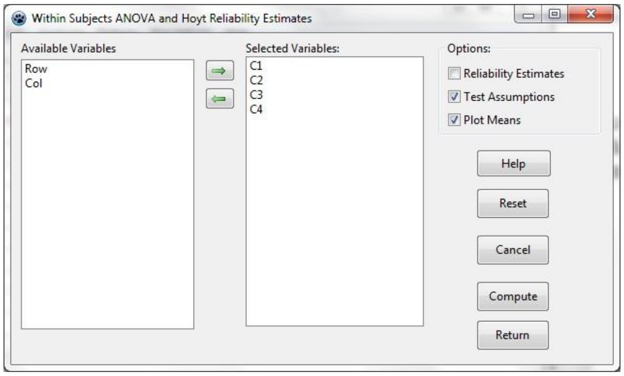


Fig. 6.2 Within Subjects ANOVA Form

Notice that the repeated measures are the columns labeled C1 through C4. You will also note that this same procedure will report intraclass reliability estimates if elected. If you now click the Compute button, you obtain the results shown below:

```
Treatments by Subjects (AxS) ANOVA Results.

Data File = C:\Users\wgmiller\LazStats\LazStatsData\ABRDATA.LAZ


-----------------------------------------------------------
SOURCE            DF        SS         MS         F  Prob. > F
-----------------------------------------------------------
SUBJECTS          11   181.000    16.455
WITHIN SUBJECTS   36  1077.000    29.917
   TREATMENTS      3   991.500   330.500   127.561     0.000
   RESIDUAL       33    85.500     2.591
-----------------------------------------------------------
TOTAL             47  1258.000    26.766
-----------------------------------------------------------


TREATMENT (COLUMN) MEANS AND STANDARD DEVIATIONS
VARIABLE  MEAN      STD.DEV.
C1          16.500     2.067
C2          11.500     2.431
C3           7.750     2.417
C4           4.250     2.864

Mean of all scores =     10.000 with standard deviation =      5.174

BOX TEST FOR HOMOGENEITY OF VARIANCE-COVARIANCE MATRIX


SAMPLE COVARIANCE MATRIX with   12 cases.


Variables
                     C1           C2           C3           C4
```

C1	4.273	2.455	1.227	1.318
C2	2.455	5.909	4.773	5.591
C3	1.227	4.773	5.841	5.432
C4	1.318	5.591	5.432	8.205

ASSUMED POP. COVARIANCE MATRIX with 12 cases.

Variables

	C1	C2	C3	C4
C1	6.057	0.693	0.693	0.693
C2	0.114	5.977	0.614	0.614
C3	0.114	0.103	5.914	0.551
C4	0.114	0.103	0.093	5.863

```
Determinant of variance-covariance matrix =       81.6
Determinant of homogeneity matrix =   1.26E003
ChiSquare =    108.149 with   8 degrees of freedom
Probability of larger chisquare = 9.66E-007
```

Treatment by Subjects (AxS Mixed Design)

We will employ the same data set used in the previous analysis. We select the AxS ANOVA option in the Comparisons sub-menu of the Statistics menu and complete the specifications on the form as show below:

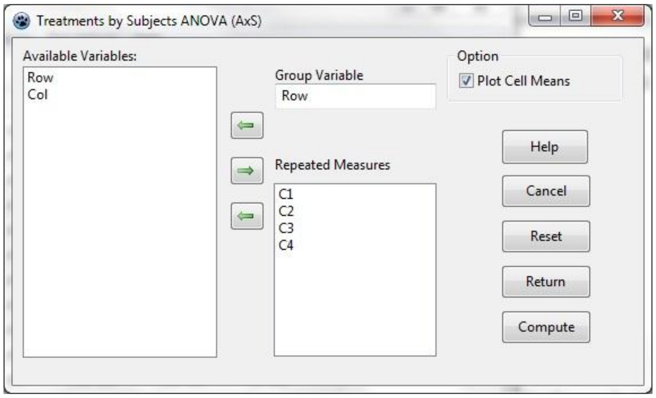


Fig. 6.3 Treatment by Subjects ANOVA Form

When the Compute button is clicked you should see these results:

```
ANOVA With One Between Subjects and One Within Subjects Treatments

------------------------------------------------------------------
Source              df      SS          MS          F          Prob.
------------------------------------------------------------------
Between             11    181.000
   Groups (A)        1     10.083     10.083      0.590       0.4602
   Subjects w.g.    10    170.917     17.092

Within Subjects     36   1077.000
   B Treatments      3    991.500    330.500    128.627       0.0000
   A X B inter.      3      8.417      2.806      1.092       0.3677
   B X S w.g.       30     77.083      2.569

TOTAL               47   1258.000
------------------------------------------------------------------
Means
TRT.   B  1   B  2   B  3   B  4   TOTAL
 A
  1   16.167 11.000  7.833  3.167  9.542
  2   16.833 12.000  7.667  5.333 10.458
TOTAL 16.500 11.500  7.750  4.250 10.000

Standard Deviations
TRT.   B  1   B  2   B  3   B  4   TOTAL
 A
  1    2.714  2.098  2.714  1.835  5.316
  2    1.329  2.828  2.338  3.445  5.099
TOTAL  2.067  2.431  2.417  2.864  5.174
```

Notice there appears to be no significant difference between the two groups of subjects but that within the groups, the first two treatment means appear to be significantly larger than the last two.

Since we elected to plot the means, we would also obtain the Fig. shown below:

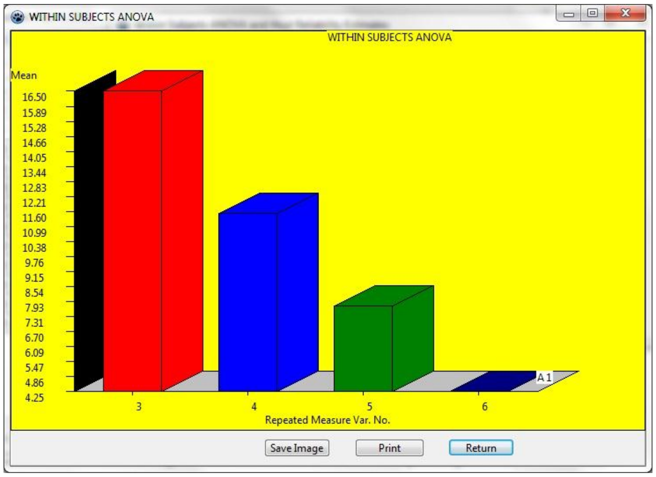


Fig. 6.4   Plot of Treatment by Subjects ANOVA Means

The graphics again demonstrate the greatest differences appear to be among the repeated measures and not the groups (A1 and A2.)

You may also have a design with two between-groups factors and repeated measures within each cell composed of subjects randomly assigned to the factor A and factor B level combinations. If you have such a design, you can employ the AxBxR Anova procedure in the LazStats package.

Two Factor Repeated Measures Analysis

Repeated measures designs have the advantage that the error terms are typically smaller that designs using independent groups of observations. This was true for the Student t-test using matched or correlated scores. On the down-side, repeated measures on the same objects pose a special problem, particularly when the objects are human subjects. The main problem is "practice" or "learning" effects that may be greater for one treatment level than another. These effects are completely confounded with the actual treatment effects. While random or counter-balanced assignment of the treatments may reduce the cumulative effects to some degree, it does not remove the effects specific to a given treatment. It is also assumed that the covariance matrices are equal among the treatment levels. Users of these designs with human subjects should be careful to minimize the practice effects. This can sometimes be done by having subjects do tasks that are similar to those in the actual experiment before beginning trials of the experiment.

In this analysis, subjects (or objects) are observed (measured) under two different treatment levels (Factors A and B levels) . For example, there might be two levels of a Factor A and three levels of a Factor B for a total of 2 x 3 = 6 treatment level combinations. Each subject would be observed 6 times in all. There must be the same subjects in each of the combinations.

The data file analyzed must consist of 4 columns of information for each observation: a variable containing an integer identification code for the subject (1..N), an integer from 1 to A for the treatment level of A, an integer from 1 to B for the treatment level of the Factor B, and a floating point variable for the observation (measurement).

A sample file (tworepeated.LAZ) was created from the example given by Quinn McNemar in his text book "Psychological Statistics", fourth edition, John Wiley and Sons, Inc., 1969, page 367. The data represent an experiment in which four subjects are observed under two levels of illumination and three levels of Albedo (Factors A and B.) The data file therefore contains 24 observations (4 x 2 x 3.) The analysis is initiated by loading the file and clicking on the "Two Within Subjects" option in the Analyses of Variance menu. The form which appears is shown below. Notice that the options have been selected to plot means of the two main effects and the interaction effects. An option has also been clicked to obtain post-hoc comparisons among the 6 means for the treatment combinations.

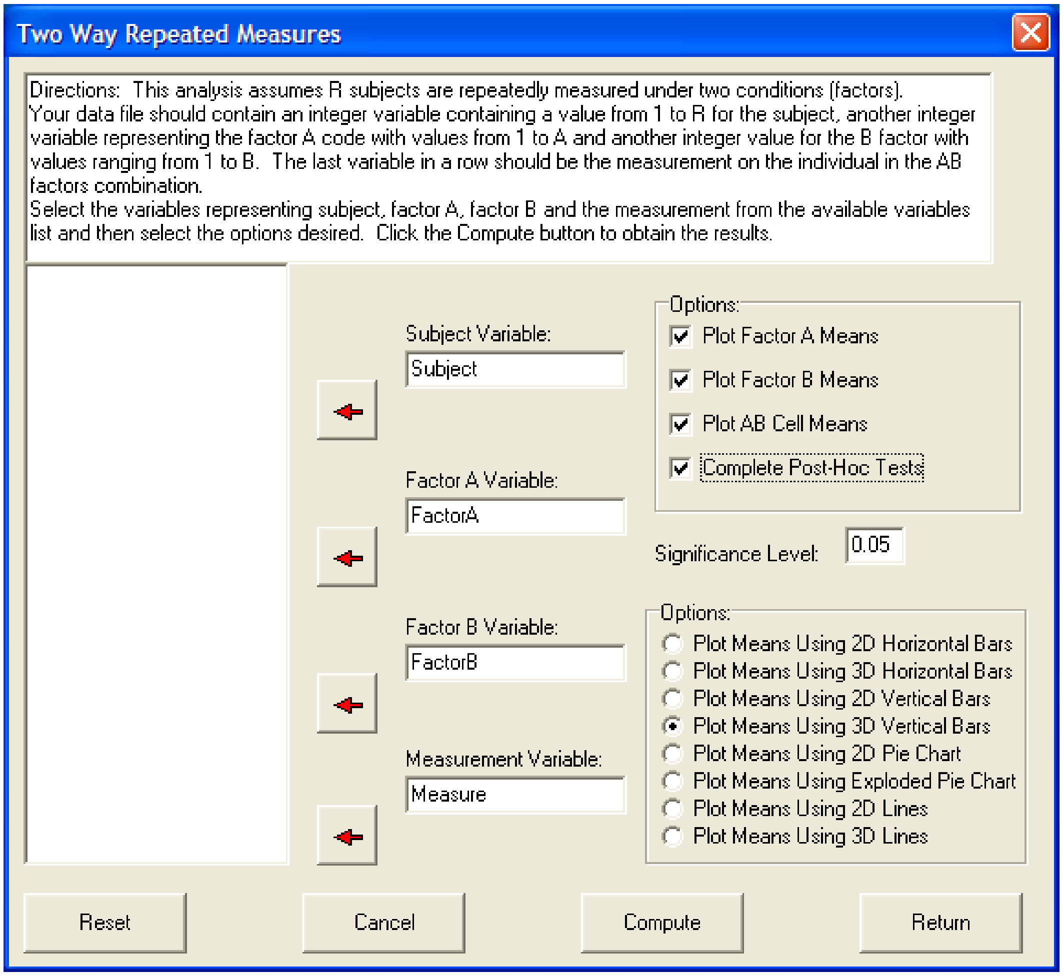


Fig. 6.5 Form for the Two-Way Repeated Measures ANOVA

When the "Compute" button is clicked the following output is obtained:

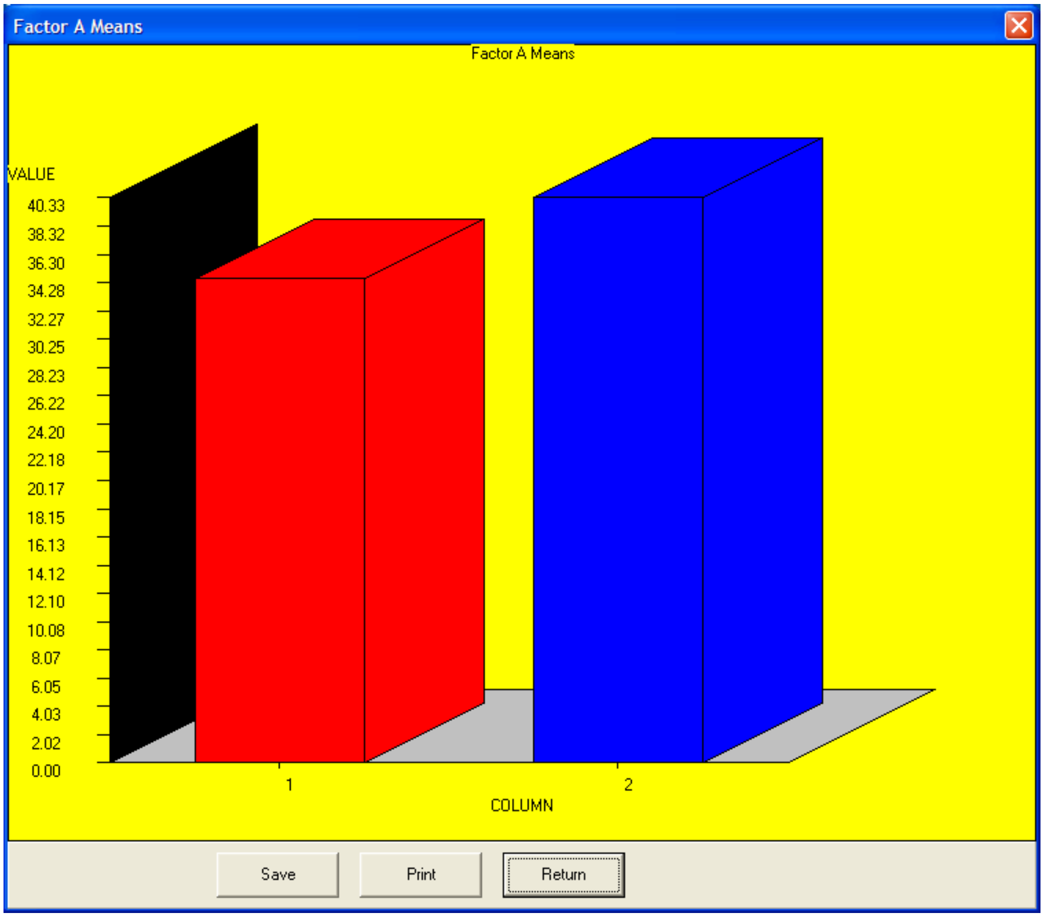


Fig. 6.6   Plot of Factor A Means in the Two-Way Repeated Measures Analysis

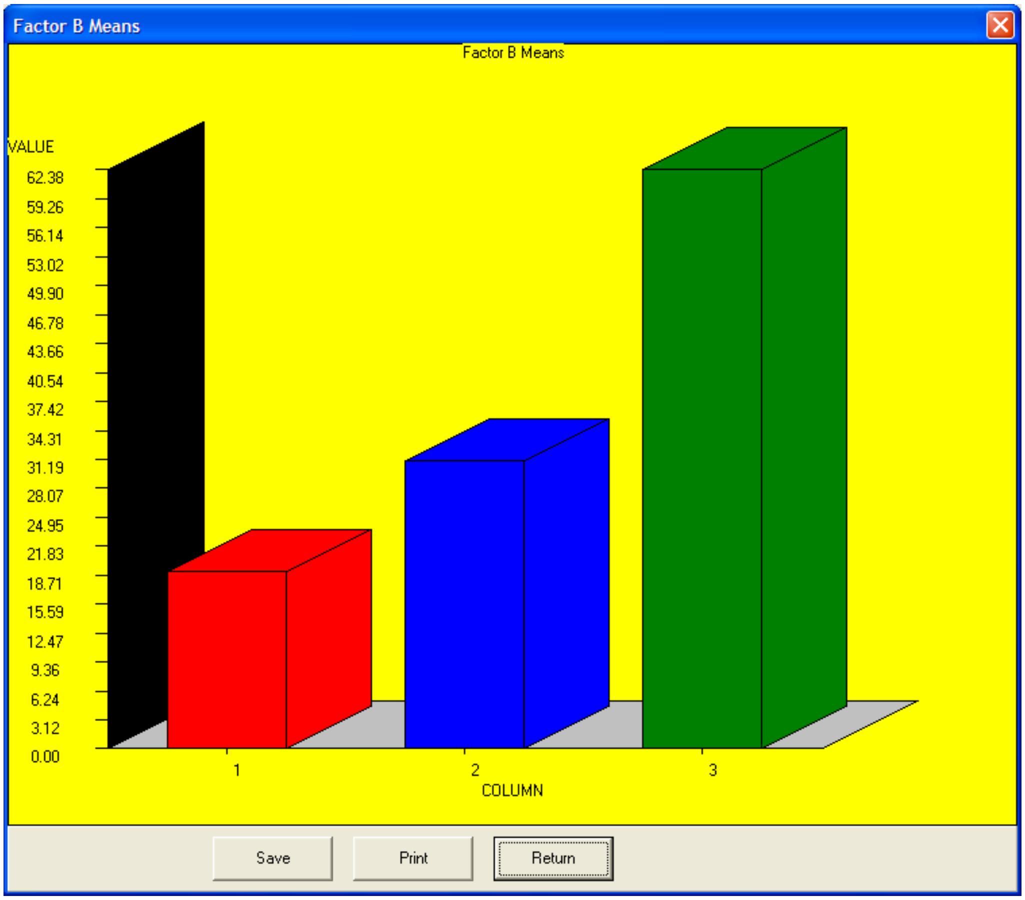


Fig. 6.7   Plot of Factor B in the Two-Way Repeated Measures Analysis

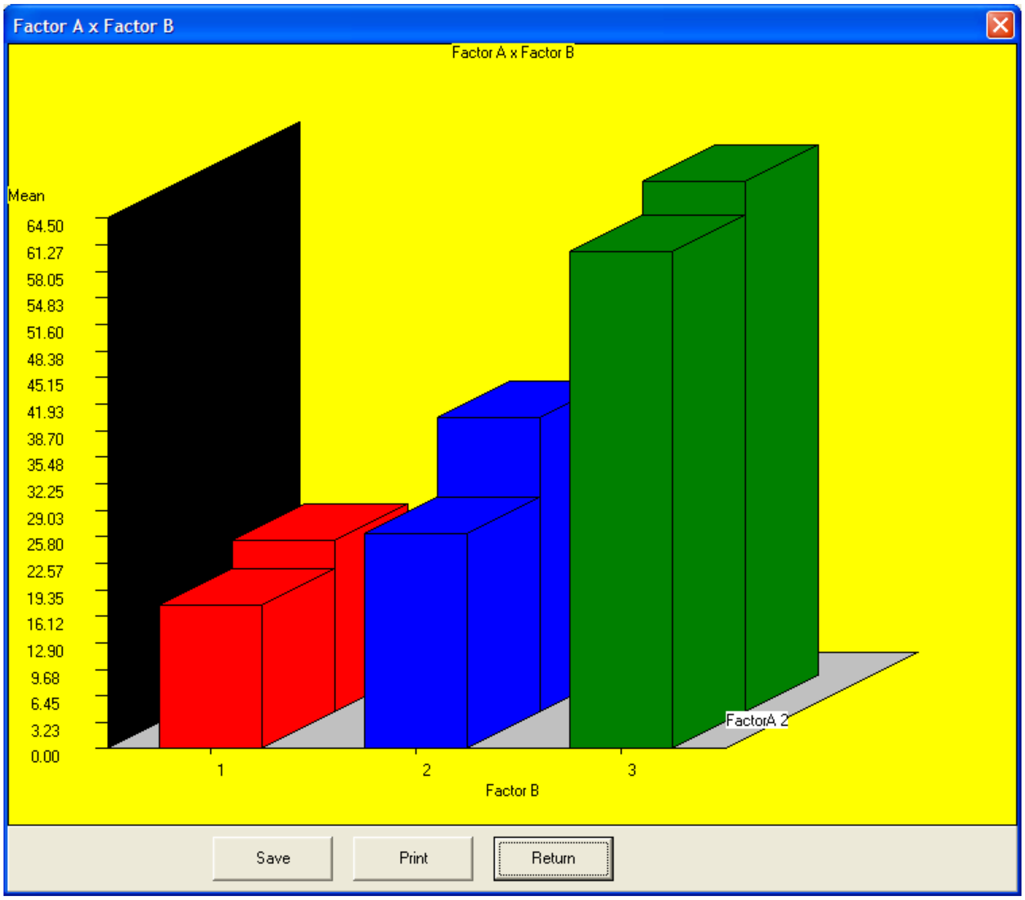


Fig. 6.8 Plot of Factor A and Factor B Interaction in the Two-Way Repeated Measures Analysis

```
-------------------------------------------------------------------
SOURCE              DF          SS           MS          F     Prob.>F
-------------------------------------------------------------------
Factor A             1       204.167      204.167      9.853    0.052
Factor B             2      8039.083     4019.542     24.994    0.001
Subjects             3      1302.833      434.278
A x B Interaction    2        46.583       23.292      0.803    0.491
A x S Interaction    3        62.167       20.722
B x S Interaction    6       964.917      160.819
A x B x S Inter.     6       174.083        29.01
-------------------------------------------------------------------
Total               23     10793.833
-------------------------------------------------------------------
Group 1 : Mean for cell A 1 and B 1 =     17.250
Group 2 : Mean for cell A 1 and B 2 =     26.000
Group 3 : Mean for cell A 1 and B 3 =     60.250
Group 4 : Mean for cell A 2 and B 1 =     20.750
Group 5 : Mean for cell A 2 and B 2 =     35.750
Group 6 : Mean for cell A 2 and B 3 =     64.500

Means for Factor A
Group 1 Mean =      34.500
Group 2 Mean =      40.333
```

```
Means for Factor B
Group 1 Mean =     19.000
Group 2 Mean =     30.875
Group 3 Mean =     62.375


---------------------------------------------------------------
             Tukey HSD Test for (Differences Between Means
                            alpha selected = 0.05
Groups      Difference  Statistic      Probability  Significant?
---------------------------------------------------------------
 1 -  2       -8.750     q =  3.249         0.3192       NO
 1 -  3      -43.000     q = 15.966         0.0004       YES
 1 -  4       -3.500     q =  1.300         0.9278       NO
 1 -  5      -18.500     q =  6.869         0.0206       YES
 1 -  6      -47.250     q = 17.544         0.0003       YES
---------------------------------------------------------------
 2 -  3      -34.250     q = 12.717         0.0009       YES
 2 -  4        5.250     q =  1.949         0.7388       NO
 2 -  5       -9.750     q =  3.620         0.2396       NO
 2 -  6      -38.500     q = 14.295         0.0006       YES
---------------------------------------------------------------
 3 -  4       39.500     q = 14.666         0.0005       YES
 3 -  5       24.500     q =  9.097         0.0052       YES
 3 -  6       -4.250     q =  1.578         0.8593       NO
---------------------------------------------------------------
 4 -  5      -15.000     q =  5.570         0.0523       NO
 4 -  6      -43.750     q = 16.244         0.0004       YES
---------------------------------------------------------------
 5 -  6      -28.750     q = 10.675         0.0023       YES
---------------------------------------------------------------


---------------------------------------------------------------
           Tukey-Kramer Test for (Differences Between Means
                     alpha selected = 0.05
Groups      Difference  Statistic      Probability  Significant?
---------------------------------------------------------------
 1 -  2      -8.750     q =  3.249         0.3192       NO
 1 -  3     -43.000     q = 15.966         0.0004       YES
 1 -  4      -3.500     q =  1.300         0.9278       NO
 1 -  5     -18.500     q =  6.869         0.0206       YES
 1 -  6     -47.250     q = 17.544         0.0003       YES
 2 -  3     -34.250     q = 12.717         0.0009       YES
 2 -  4       5.250     q =  1.949         0.7388       NO
 2 -  5      -9.750     q =  3.620         0.2396       NO
 2 -  6     -38.500     q = 14.295         0.0006       YES
 3 -  4      39.500     q = 14.666         0.0005       YES
 3 -  5      24.500     q =  9.097         0.0052       YES
 3 -  6      -4.250     q =  1.578         0.8593       NO
 4 -  5     -15.000     q =  5.570         0.0523       NO
 4 -  6     -43.750     q = 16.244         0.0004       YES
 5 -  6     -28.750     q = 10.675         0.0023       YES
---------------------------------------------------------------

---------------------------------------------------------------
           Tukey B Test for (Contrasts on Ordered Means
                          alpha selected = 0.05
```

```
---------------------------------------------------------------

Groups     Difference  Statistic   d.f.     Prob.>value  Significant?
  1 -   4     -3.500       1.300     2,   6  0.661        NO
  1 -   2     -8.750       3.249     3,   6  0.226        NO
  1 -   5    -18.500       6.869     4,   6  0.016        YES
  1 -   3    -43.000      15.966     5,   6  0.000        YES
  1 -   6    -47.250      17.544     6,   6  0.000        YES
  4 -   2     -5.250       1.949     2,   6  0.478        NO
  4 -   5    -15.000       5.570     3,   6  0.035        YES
  4 -   3    -39.500      14.666     4,   6  0.000        YES
  4 -   6    -43.750      16.244     5,   6  0.000        YES
  2 -   5     -9.750       3.620     2,   6  0.141        NO
  2 -   3    -34.250      12.717     3,   6  0.001        YES
  2 -   6    -38.500      14.295     4,   6  0.000        YES
  5 -   3    -24.500       9.097     2,   6  0.003        YES
  5 -   6    -28.750      10.675     3,   6  0.002        YES
  3 -   6     -4.250       1.578     2,   6  0.583        NO

----------------------------------------------------------------
                 Scheffe contrasts among pairs of means.
                            alpha selected = 0.05
Group vs Group  Difference   Scheffe    Critical  Significant?
                             Statistic  Value
----------------------------------------------------------------
 1         2          -8.75     2.30      3.723     NO
 1         3         -43.00    11.29      3.723     YES
 1         4          -3.50     0.92      3.723     NO
 1         5         -18.50     4.86      3.723     YES
 1         6         -47.25    12.41      3.723     YES
----------------------------------------------------------------
 2         3         -34.25     8.99      3.723     YES
 2         4           5.25     1.38      3.723     NO
 2         5          -9.75     2.56      3.723     NO
 2         6         -38.50    10.11      3.723     YES
----------------------------------------------------------------
 3         4          39.50    10.37      3.723     YES
 3         5          24.50     6.43      3.723     YES
 3         6          -4.25     1.12      3.723     NO
----------------------------------------------------------------
 4         5         -15.00     3.94      3.723     YES
 4         6         -43.75    11.49      3.723     YES
----------------------------------------------------------------
 5         6         -28.75     7.55      3.723     YES
----------------------------------------------------------------

----------------------------------------------------------------------
            Neuman-Keuls Test for (Contrasts on Ordered Means
                            alpha selected = 0.05

Group     Mean
  1      17.250
  4      20.750
  2      26.000
  5      35.750
  3      60.250
  6      64.500

Groups      Difference  Statistic      d.f.   Probability  Significant?
```

```
---------------------------------------------------------------------
 1 -  4      -3.500      q =  1.300    2    6     0.3935      NO
 1 -  2      -8.750      q =  3.249    3    6     0.1323      NO
 1 -  5     -18.500      q =  6.869    4    6     0.0112      YES
 1 -  3     -43.000      q = 15.966    5    6     0.0003      YES
 1 -  6     -47.250      q = 17.544    6    6     0.0003      YES
 4 -  2      -5.250      q =  1.949    2    6     0.2174      NO
 4 -  5     -15.000      q =  5.570    3    6     0.0180      YES
 4 -  3     -39.500      q = 14.666    4    6     0.0002      YES
 4 -  6     -43.750      q = 16.244    5    6     0.0003      YES
 2 -  5      -9.750      q =  3.620    2    6     0.0430      YES
 2 -  3     -34.250      q = 12.717    3    6     0.0004      YES
 2 -  6     -38.500      q = 14.295    4    6     0.0003      YES
 5 -  3     -24.500      q =  9.097    2    6     0.0008      YES
 5 -  6     -28.750      q = 10.675    3    6     0.0008      YES
 3 -  6      -4.250      q =  1.578    2    6     0.3070      NO
---------------------------------------------------------------------
```

The above results reflect possible significance for the main effects of Factors A and B but not for the interaction. The F ratio of the Factor A is obtained by dividing the mean square for Factor A by the mean square for interaction of subjects with Factor A. In a similar manner, the F ratio for Factor B is the ratio of the mean square for Factor B to the mean square of the interaction of Factor B with subjects. Finally, the F ratio for the interaction of Factor A with Factor B uses the triple interaction of A with B with Subjects as the denominator.

Between 5 or 6 of the post-hoc comparisons were not significant among the 15 possible comparisons among means using the 0.05 level for rejection of the hypothesis of no difference.

Nested Factors Analysis Of Variance Design

Shown below is an example of a nested analysis using the file ABNested.LAZ. When you select this analysis, you see the dialog below:

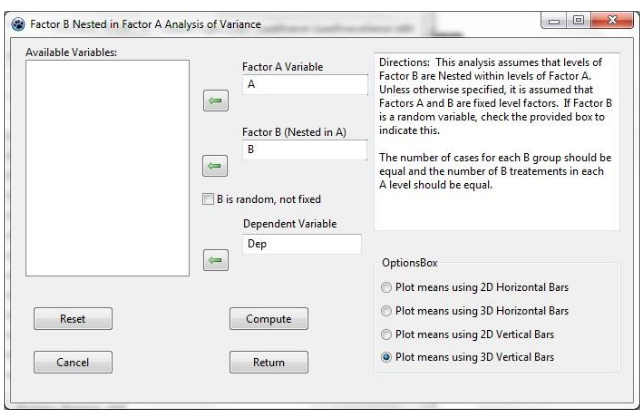


Fig. 6.9 The Nested ANOVA Form

The results are shown below:

```
Nested ANOVA by Bill Miller
File Analyzed = C:\lazarus\Projects\LazStats\LazStatsData\ABNested.LAZ
CELL MEANS
A LEVEL     BLEVEL         MEAN          STD.DEV.
    1          1            2.667         1.528
    1          2            3.333         1.528
    1          3            4.000         1.732
```

```
    2         4             3.667          1.528
    2         5             4.000          1.000
    2         6             5.000          1.000
    3         7             3.667          1.155
    3         8             5.000          1.000
    3         9             6.333          0.577

A MARGIN MEANS
A LEVEL       MEAN             STD.DEV.
    1          3.333            1.500
    2          4.222            1.202
    3          5.000            1.414

GRAND MEAN =       4.185

ANOVA TABLE
SOURCE     D.F.        SS         MS         F          PROB.
A            2       12.519     6.259     3.841      0.041
B(W)         6       16.222     2.704     1.659      0.189
w.cells     18       29.333     1.630
Total       26       58.074
```

Of course, if you elect to plot the means, additional graphical output is included.

A, B and C Factors with B Nested in A

Shown below is the dialog for this ANOVA design and the results of analyzing the file ABCNested.LAZ:

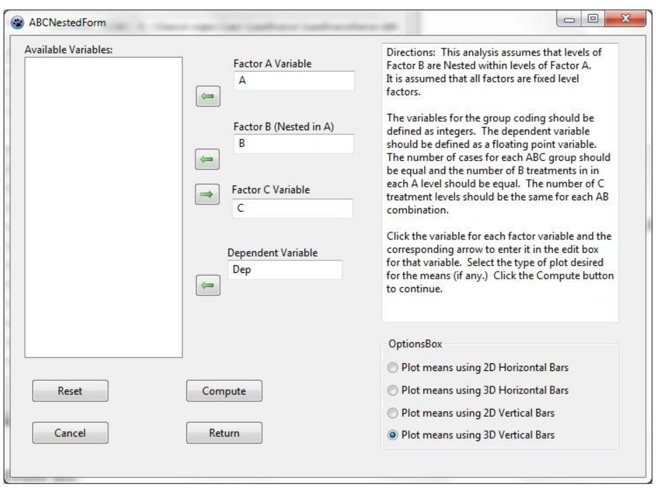


Fig. 6.10   Three Factor Nested ANOVA

The results are:

```
Nested ANOVA by Bill Miller
File Analyzed = C:\lazarus\Projects\LazStats\LazStatsData\ABCNested.LAZ

CELL MEANS
A LEVEL     BLEVEL       CLEVEL        MEAN       STD.DEV.
    1           1            1        2.6667        1.5275
    1           1            2        3.3333        1.1547
    1           2            1        3.3333        1.5275
    1           2            2        3.6667        2.0817
```

1	3	1	4.0000	1.7321
1	3	2	5.0000	1.7321
2	4	1	3.6667	1.5275
2	4	2	4.6667	1.5275
2	5	1	4.0000	1.0000
2	5	2	4.6667	0.5774
2	6	1	5.0000	1.0000
2	6	2	3.0000	1.0000
3	7	1	3.6667	1.1547
3	7	2	2.6667	1.1547
3	8	1	5.0000	1.0000
3	8	2	6.0000	1.0000
3	9	1	6.6667	1.1547
3	9	2	6.3333	0.5774

A MARGIN MEANS

A LEVEL	MEAN	STD.DEV.
1	3.667	1.572
2	4.167	1.200
3	5.056	1.731

B MARGIN MEANS

B LEVEL	MEAN	STD.DEV.
1	3.000	1.265
2	3.500	1.643
3	4.500	1.643
4	4.167	1.472
5	4.333	0.816
6	4.000	1.414
7	3.167	1.169
8	5.500	1.049
9	6.500	0.837

C MARGIN MEANS

C LEVEL	MEAN	STD.DEV.
1	4.222	1.577
2	4.370	1.644

AB MARGIN MEANS

A LEVEL	B LEVEL	MEAN	STD.DEV.
1	1	3.000	1.265
1	2	3.500	1.643
1	3	4.500	1.643
2	4	4.167	1.472
2	5	4.333	0.816
2	6	4.000	1.414
3	7	3.167	1.169
3	8	5.500	1.049
3	9	6.500	0.837

AC MARGIN MEANS

A LEVEL	C LEVEL	MEAN	STD.DEV.
1	1	0.000	1.500
1	2	0.000	1.658
2	1	0.000	1.202
2	2	0.000	1.269
3	1	0.000	1.616
3	2	0.000	1.936

```
GRAND MEAN =      4.296

ANOVA TABLE
SOURCE     D.F.        SS         MS         F          PROB.
A            2      17.815     8.907     5.203     0.010
B(A)         6      42.444     7.074     4.132     0.003
C            1       0.296     0.296     0.173     0.680
AxC          2       1.815     0.907     0.530     0.593
B(A)xC       6      11.556     1.926     1.125     0.368
w.cells     36      61.630     1.712
Total       53     135.259
```

Latin and Greco-Latin Square Designs

Example in Education Using a Latin Square

Assume you are interested in the achievement of students under three methods of instruction for a required course in biology (self, computer, and classroom), interested in differences of these instruction modes for three colleges within a university (agriculture, education, engineering) and three types of students (in-state, out-of-state, out-of-country). We could use a completely balanced 3-way analysis of variance design with Factor A = instructional mode, Factor B = College and Factor C = type of student. There would be 27 experimental units (samples of subjects) in this design. On the other hand we might employ the following design:

	FACTOR A (Instruction)		
	Self	Computer	Classroom
FACTOR B (College)			
Agriculture	C2	C1	C3
Education	C1	C3	C2
Engineering	C3	C2	C1

In this design C1 is the in-state student unit, C2 is the out-of-state student unit and C3 is the out-of-country student unit. There are only 9 units in this design as contrasted with 27 units in the completely balanced design. Note that each type of student receives each type of instruction. Also note however that, within a college, students of each type do NOT receive each type of instruction. We will have to assume that the interaction of college and type of instruction, the interaction of college and type of student, the interaction of type of instruction and type of student and the triple interaction of College, instruction and student are small or do not exist. We are primarily interested in the main effects, that is, differences among student types, types of instruction and colleges on the achievement scores obtained in the biology course. We might use Plan 1 described below.

Plan 1 by B.J. Winer

We have prepared an example file for you to analyze with LazStats. Open the file labeled LatinSqr.LAZ in your set of sample data files. We have entered four cases for each unit in our design for instructional mode, college and home residence. Once you have loaded the file, select the Latin squares designs option under the sub-menu for comparisons under the Analyses menu. You should see the form below for selecting the Plan 1 analysis.

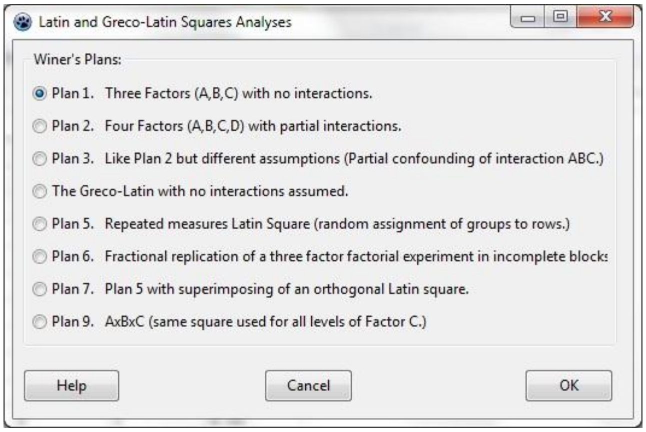


Fig. 6.11 Latin and Greaco-Latin Squares Form

When you have selected Plan 1 for the analysis, click the OK button to continue. You will then see the form below for entering the specifications for your analysis. We have entered the variables for factors A, B and C and entered the number of cases for each unit:

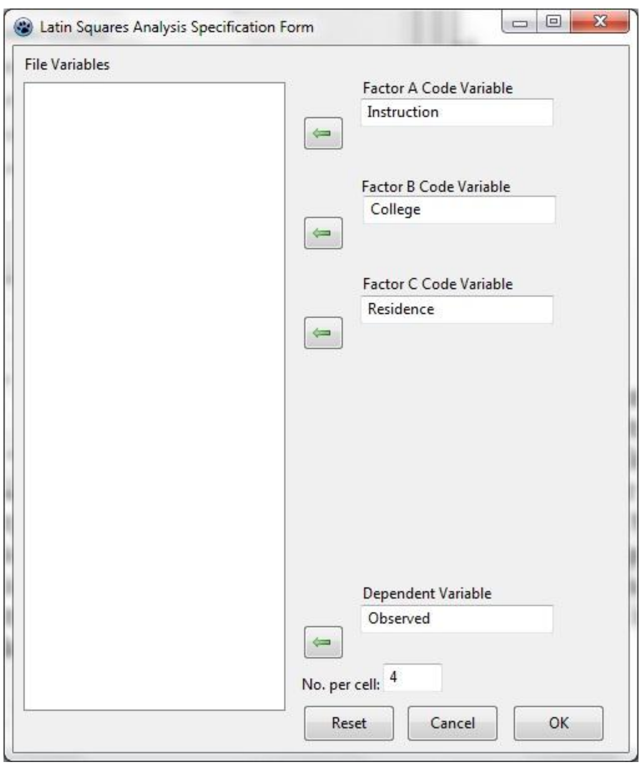


Fig. 6.12 Latin Squares Analysis Dialog

We have completed the entry of our variables and the number of cases and are ready to continue. When you press the OK button, the following results are presented on the output page:

```
Latin Square Analysis Plan 1 Results

-----------------------------------------------------------
Source           SS          DF         MS          F       Prob.>F
-----------------------------------------------------------
Factor A      92.389          2     46.194     12.535      0.000
Factor B      40.222          2     20.111      5.457      0.010
```

```
Factor C    198.722         2    99.361    26.962     0.000
Residual     33.389         2    16.694     4.530     0.020
Within       99.500        27     3.685
Total       464.222        35
-----------------------------------------------------------

Experimental Design
------------------------------
Instruction   1    2    3
------------------------------
   College
     1       C2   C3   C1
     2       C3   C1   C2
     3       C1   C2   C3
------------------------------


Cell means and totals
--------------------------------------------------
Instruction     1         2         3        Total
--------------------------------------------------
   College
     1        2.750    10.750     3.500     5.667
     2        8.250     2.250     1.250     3.917
     3        1.500     1.500     2.250     1.750
Total         4.167     4.833     2.333     3.778
--------------------------------------------------

--------------------------------------------------
 Residence      1         2         3        Total
--------------------------------------------------
              2.417     1.833     7.083     3.778
--------------------------------------------------
```

A partial test of the interaction effects can be made by the ratio of the MS for residual to the MS within cells. In our example, it appears that our assumptions of no interaction effects may be in error. In this case, the main effects may be confounded by interactions among the factors. The results may never the less suggest differences do exist and we should complete another balanced experiment to determine the interaction effects.

Plan 2

We have included the file “LatinSqr2.LAZ” as an example for analysis. Load the file in the grid and select the Latin Square Analyses, Plan 2 design. The form below shows the entry of the variables and the sample size for the analysis:

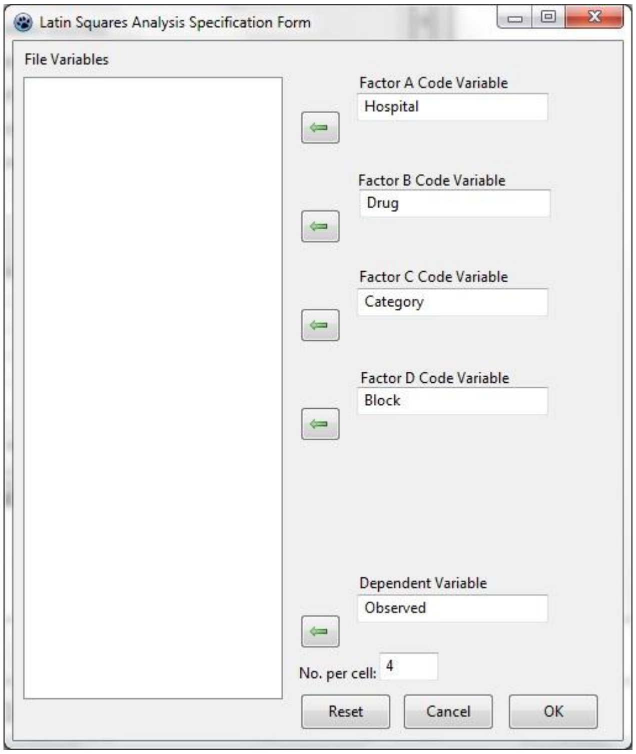


Fig. 6.13 Four Factor Latin Square Design Form

When you click the OK button, you will see the following results:

Latin Square Analysis Plan 2 Results

Source	SS	DF	MS	F	Prob.>F
Factor A	148.028	2	74.014	20.084	0.000
Factor B	5.444	2	2.722	0.739	0.483
Factor C	66.694	2	33.347	9.049	0.000
Factor D	18.000	1	18.000	4.884	0.031
A x D	36.750	2	18.375	4.986	0.010
B x D	75.000	2	37.500	10.176	0.000
C x D	330.750	2	165.375	44.876	0.000
Residual	66.778	4	16.694	4.530	0.003
Within	199.000	54	3.685		
Total	946.444	71			

Experimental Design for block 1

Drug	1	2	3
Hospital			
1	C2	C3	C1
2	C3	C1	C2
3	C1	C2	C3

Experimental Design for block 2

Drug	1	2	3
Hospital			
1	C2	C3	C1
2	C3	C1	C2
3	C1	C2	C3

BLOCK 1

Cell means and totals

Drug	1	2	3	Total
Hospital				
1	2.750	10.750	3.500	5.667
2	8.250	2.250	1.250	3.917
3	1.500	1.500	2.250	1.750
Total	4.167	4.833	2.333	4.278

BLOCK 2

Cell means and totals

Drug	1	2	3	Total
Hospital				
1	9.250	2.250	3.250	4.917
2	3.750	4.500	11.750	6.667
3	2.500	3.250	2.500	2.750
Total	5.167	3.333	5.833	4.278

Category	1	2	3	Total
	2.917	4.958	4.958	4.278

Notice that the interactions with Factor D are obtained. The residual however indicates that some of the other interactions confounded with the main factors may be significant and, again, we do not know the portion of the differences among the main effects that are potentially due to interactions among A, B, and C.

Plan 3 Latin Squares Design

The file “LatinSqr3.LAZ” contains an example of data for the Plan 3 analysis. Following the previous plans, we show below the specifications for the analysis and results from analyzing this data:

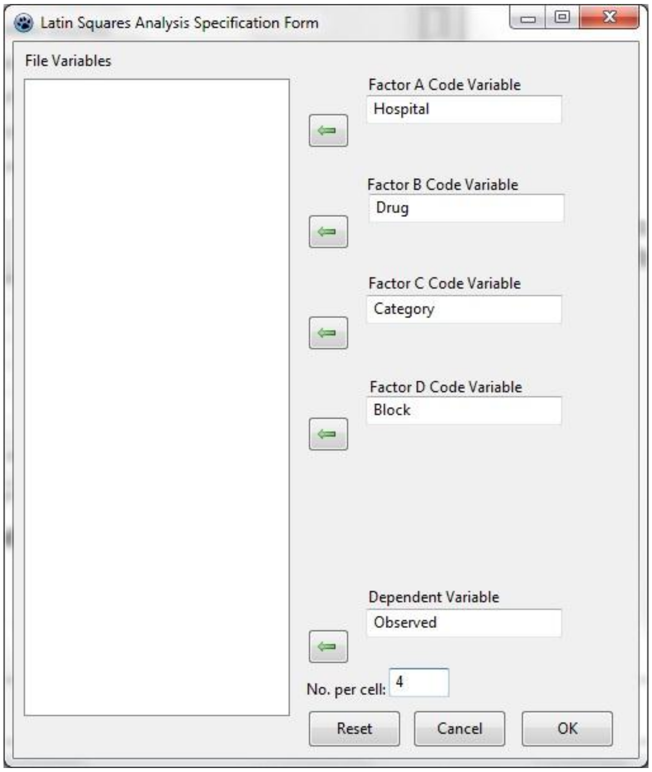


Fig. 6.14 Another Latin Square (Plan 3) Dialog Form

Latin Square Analysis Plan 3 Results

Source	SS	DF	MS	F	Prob.>F
Factor A	26.963	2	13.481	3.785	0.027
Factor B	220.130	2	110.065	30.902	0.000
Factor C	213.574	2	106.787	29.982	0.000
Factor D	19.185	2	9.593	2.693	0.074
A x B	49.148	4	12.287	3.450	0.012
A x C	375.037	4	93.759	26.324	0.000
B x C	78.370	4	19.593	5.501	0.001
A x B x C	118.500	6	19.750	5.545	0.000
Within	288.500	81	3.562		
Total	1389.407	107			

Experimental Design for block 1

Drug	1	2	3
Hospital			
1	C1	C2	C3
2	C2	C3	C1
3	C3	C1	C2

Experimental Design for block 2

Drug	1	2	3

```
  Hospital
     1       C2   C3   C1
     2       C3   C1   C2
     3       C1   C2   C3
------------------------------

Experimental Design for block 3
------------------------------
      Drug   1    2    3
------------------------------
  Hospital
     1       C3   C1   C2
     2       C1   C2   C3
     3       C2   C3   C1
------------------------------

BLOCK 1

Cell means and totals
--------------------------------------------------
      Drug     1         2         3         Total
--------------------------------------------------
  Hospital
     1        2.750     1.250     1.500     1.833
     2        3.250     4.500     2.500     3.417
     3       10.250     8.250     2.250     6.917
Total         5.417     4.667     2.083     4.074
--------------------------------------------------

BLOCK 2

Cell means and totals
--------------------------------------------------
      Drug     1         2         3         Total
--------------------------------------------------
  Hospital
     1       10.750     8.250     2.250     7.083
     2        9.250    11.750     3.250     8.083
     3        3.500     1.750     1.500     2.250
Total         7.833     7.250     2.333     4.074
--------------------------------------------------

BLOCK 3

Cell means and totals
--------------------------------------------------
      Drug     1         2         3         Total
--------------------------------------------------
  Hospital
     1        3.500     2.250     1.500     2.417
     2        2.250     3.750     2.500     2.833
     3        2.750     1.250     1.500     1.833
Total         2.833     2.417     1.833     4.074
--------------------------------------------------

Means for each variable

--------------------------------------------------
  Hospital     1         2         3         Total
```

```
---------------------------------------------------
              3.778      4.778      3.667      4.074
---------------------------------------------------

---------------------------------------------------
      Drug      1          2          3         Total
---------------------------------------------------
              5.361      4.778      2.083      4.074
---------------------------------------------------

---------------------------------------------------
  Category      1          2          3         Total
---------------------------------------------------
              4.056      5.806      2.361      4.074
---------------------------------------------------

---------------------------------------------------
     Block      1          2          3         Total
---------------------------------------------------
              4.500      4.222      3.500      4.074
---------------------------------------------------
```

Here, the main effect of factor D is partially confounded with the ABC interaction.

Analysis of Greco-Latin Squares

The file labeled "LatinGreco.LAZ" contains sample data for a Greco-Latin design analysis.

The specifications for the analysis are entered as:

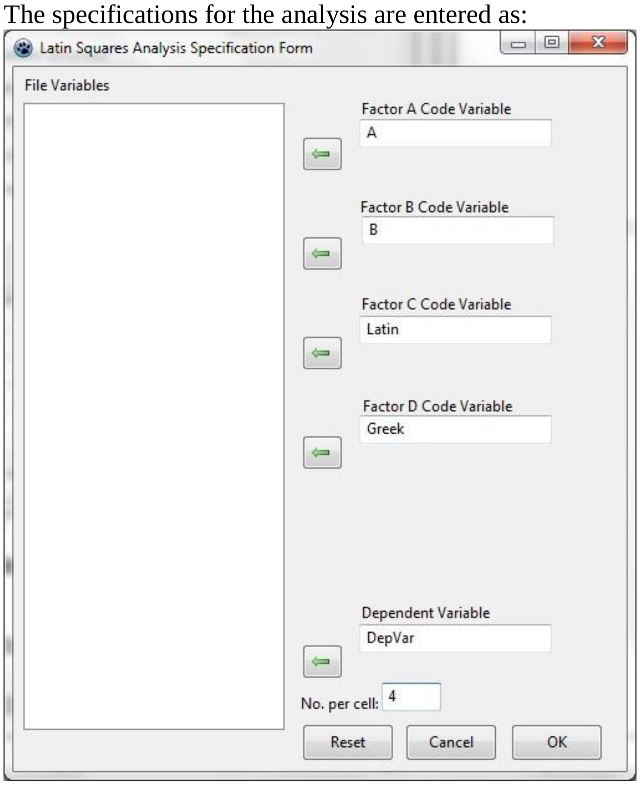


Fig. 6.15 Latin Square Design Form

The results are obtained as:

Greco-Latin Square Analysis (No Interactions)

Source	SS	DF	MS	F	Prob.>F
Factor A	64.889	2	32.444	9.733	0.001
Factor B	64.889	2	32.444	9.733	0.001
Latin Sqr.	24.889	2	12.444	3.733	0.037
Greek Sqr.	22.222	2	11.111	3.333	0.051
Residual	-	-	-	-	-
Within	90.000	27	3.333		
Total	266.889	35			

Experimental Design for Latin Square

B	1	2	3
A			
1	C1	C2	C3
2	C2	C3	C1
3	C3	C1	C2

Experimental Design for Greek Square

B	1	2	3
A			
1	C1	C2	C3
2	C3	C1	C2
3	C2	C3	C1

Cell means and totals

B	1	2	3	Total
A				
1	4.000	6.000	7.000	5.667
2	6.000	12.000	8.000	8.667
3	7.000	8.000	10.000	8.333
Total	5.667	8.667	8.333	7.556

Means for each variable

A	1	2	3	Total
	5.667	8.667	8.333	7.556

B	1	2	3	Total
	5.667	8.667	8.333	7.556

Latin	1	2	3	Total
	6.667	7.333	8.667	7.556

Greek	1	2	3	Total
	8.667	7.000	7.000	7.556

Notice that in the case of 3 levels that the residual degrees of freedom are 0 hence no term is shown for the residual in this example. For more than 3 levels the test of the residuals provides a partial check on the assumptions of negligible interactions. The residual is sometimes combined with the within cell variance to provide an over-all estimate of variation due to experimental error.

Plan 5 Latin Square Design

The specifications for the analysis of the sample file "LatinPlan5.LAZ" is shown below:

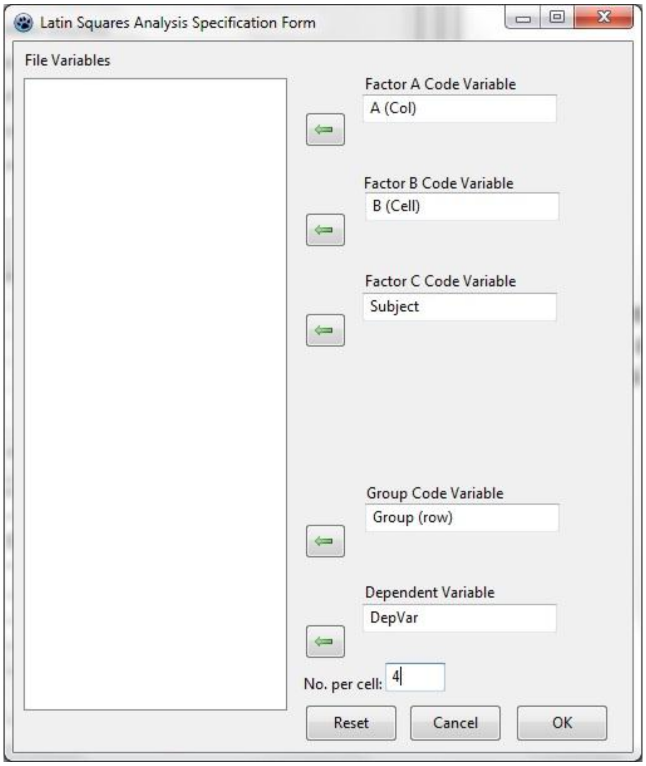


Fig. 6.16 Latin Square Plan 5 Form

If you examine the sample file, you will notice that the subject Identification numbers (1,2,3,4) for the subjects in each group are the same even though the subjects in each group are different from group to group. The same ID is used in each group because they become "subscripts" for several arrays in the program. The results for our sample data are shown below:

```
Greco-Latin Square Analysis (No Interactions)

-----------------------------------------------------------
Source          SS         DF        MS         F      Prob.>F
-----------------------------------------------------------
Factor A      64.889        2     32.444     9.733      0.001
Factor B      64.889        2     32.444     9.733      0.001
Latin Sqr.    24.889        2     12.444     3.733      0.037
Greek Sqr.    22.222        2     11.111     3.333      0.051
Residual       -         -          -          -          -
Within        90.000       27      3.333
Total        266.889       35
-----------------------------------------------------------

Experimental Design for Latin Square
------------------------------
         B   1    2    3
------------------------------
         A
     1       C1   C2   C3
     2       C2   C3   C1
     3       C3   C1   C2
------------------------------

Experimental Design for Greek Square
```

B	1	2	3
A			
1	C1	C2	C3
2	C3	C1	C2
3	C2	C3	C1

Cell means and totals

B	1	2	3	Total
A				
1	4.000	6.000	7.000	5.667
2	6.000	12.000	8.000	8.667
3	7.000	8.000	10.000	8.333
Total	5.667	8.667	8.333	7.556

Means for each variable

A	1	2	3	Total
	5.667	8.667	8.333	7.556

B	1	2	3	Total
	5.667	8.667	8.333	7.556

Latin	1	2	3	Total
	6.667	7.333	8.667	7.556

Greek	1	2	3	Total
	8.667	7.000	7.000	7.556

Sums for ANOVA Analysis

Group (rows) times A Factor (columns) sums with 36 cases.

Variables

	1	2	3	Total
1	14.000	19.000	18.000	51.000
2	15.000	18.000	16.000	49.000
3	14.000	21.000	18.000	53.000
Total	43.000	58.000	52.000	153.000

Group (rows) times B (cells Factor) sums with 36 cases.

Variables

	1	2	3	Total
1	19.000	18.000	14.000	51.000
2	15.000	18.000	16.000	49.000
3	18.000	14.000	21.000	53.000
Total	52.000	50.000	51.000	153.000

```
Groups (rows) times Subjects (columns) matrix with   36 cases.


Variables
                     1            2            3            4         Total
        1      13.000       11.000       13.000       14.000       51.000
        2      10.000       14.000       10.000       15.000       49.000
        3      13.000        9.000       17.000       14.000       53.000
    Total      36.000       34.000       40.000       43.000      153.000

Latin Squares Repeated Analysis Plan 5 (Partial Interactions)

-----------------------------------------------------------
Source          SS        DF        MS        F      Prob.>F
-----------------------------------------------------------
Betw.Subj.   20.083        11
 Groups       0.667         2     0.333     0.155     0.859
 Subj.w.g.   19.417         9     2.157

Within Sub   36.667        24
 Factor A     9.500         2     4.750     3.310     0.060
 Factor B     0.167         2     0.083     0.058     0.944
 Factor AB    1.167         2     0.583     0.406     0.672
 Error w.    25.833        18     1.435
Total        56.750        35
-----------------------------------------------------------

Experimental Design for Latin Square
------------------------------
   A (Col)   1    2    3
------------------------------
Group (row)
     1       B3   B1   B2
     2       B1   B2   B3
     3       B2   B3   B1
------------------------------

Cell means and totals
--------------------------------------------------
   A (Col)     1         2         3        Total
--------------------------------------------------
Group (row)
     1        3.500     4.750     4.500     4.250
     2        3.750     4.500     4.000     4.083
     3        3.500     5.250     4.500     4.417
Total         3.583     4.833     4.333     4.250
--------------------------------------------------

Means for each variable

--------------------------------------------------
   A (Col)     1         2         3        Total
--------------------------------------------------
              4.333     4.167     4.250     4.250
--------------------------------------------------

--------------------------------------------------
  B (Cell)     1         2         3        Total
--------------------------------------------------
              4.250     4.083     4.417     4.250
--------------------------------------------------

--------------------------------------------------
Group (row)     1         2         3        Total
--------------------------------------------------
              4.250     4.083     4.417     4.250
--------------------------------------------------
```

Plan 6 Latin Squares Design

LatinPlan6.LAZ is the name of a sample file which you can analyze with the Plan 6 option of the Latin squares analysis procedure. Shown below is the specification form for the analysis of the data in that file:

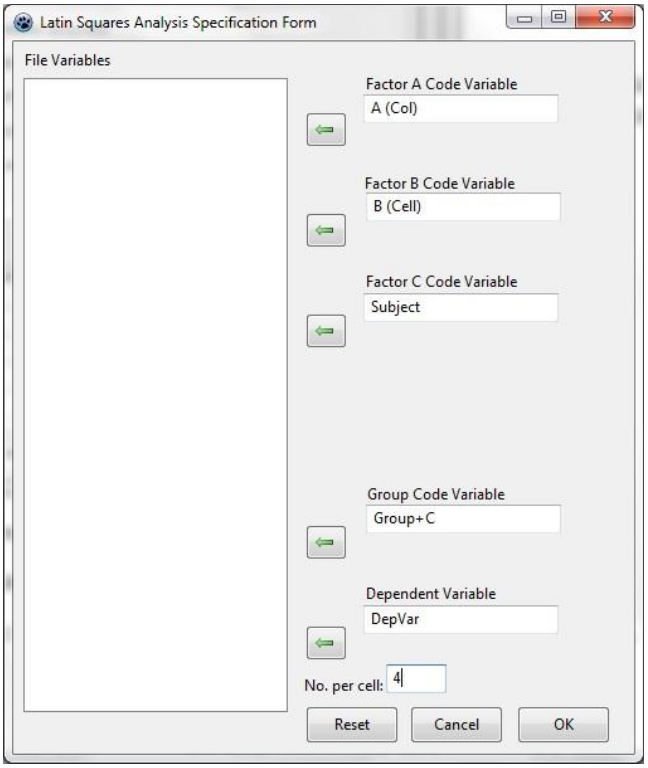


Fig. 6.17   Latin Square Plan 6 Form

The results obtained when you click the OK button are shown below:

```
Latin Squares Repeated Analysis Plan 6

Sums for ANOVA Analysis

Group - C (rows) times A Factor (columns) sums with   36 cases.

Variables
                    1            2            3          Total
        1      23.000       16.000       22.000       61.000
        2      22.000       14.000       18.000       54.000
        3      24.000       21.000       21.000       66.000
    Total      69.000       51.000       61.000      181.000


Group - C (rows) times B (cells Factor) sums with   36 cases.

Variables
                    1            2            3          Total
        1      16.000       22.000       23.000       61.000
        2      22.000       14.000       18.000       54.000
        3      21.000       24.000       21.000       66.000
    Total      59.000       60.000       62.000      181.000

Group - C (rows) times Subjects (columns) matrix with   36 cases.

Variables
                    1            2            3            4          Total
        1      16.000       14.000       13.000       18.000       61.000
        2      12.000       13.000       14.000       15.000       54.000
        3      18.000       19.000       11.000       18.000       66.000
    Total      46.000       46.000       38.000       51.000      181.000

Latin Squares Repeated Analysis Plan 6

-----------------------------------------------------------
```

Source	SS	DF	MS	F	Prob.>F
Betw.Subj.	26.306	11			
Factor C	6.056	2	3.028	1.346	0.308
Subj.w.g.	20.250	9	2.250		
Within Sub	70.667	24			
Factor A	13.556	2	6.778	2.259	0.133
Factor B	0.389	2	0.194	0.065	0.937
Residual	2.722	2	1.361	0.454	0.642
Error w.	54.000	18	3.000		
Total	96.972	35			

Experimental Design for Latin Square

G	C	A (Col) 1	2	3
1	1	B3	B1	B2
2	2	B1	B2	B3
3	3	B2	B3	B1

Cell means and totals

A (Col)	1	2	3	Total
Group+C				
1	5.750	4.000	5.500	5.083
2	5.500	3.500	4.500	4.500
3	6.000	5.250	5.250	5.500
Total	5.750	4.250	5.083	5.028

Means for each variable

A (Col)	1	2	3	Total
	4.917	5.000	5.167	5.028

B (Cell)	1	2	3	Total
	5.083	4.500	5.500	5.028

Group+C	1	2	3	Total
	5.083	4.500	5.500	5.028

Plan 7 for Latin Squares

Shown below is the specification for analysis of the sample data file labeled LatinPlan7.LAZ and the results of the analysis:

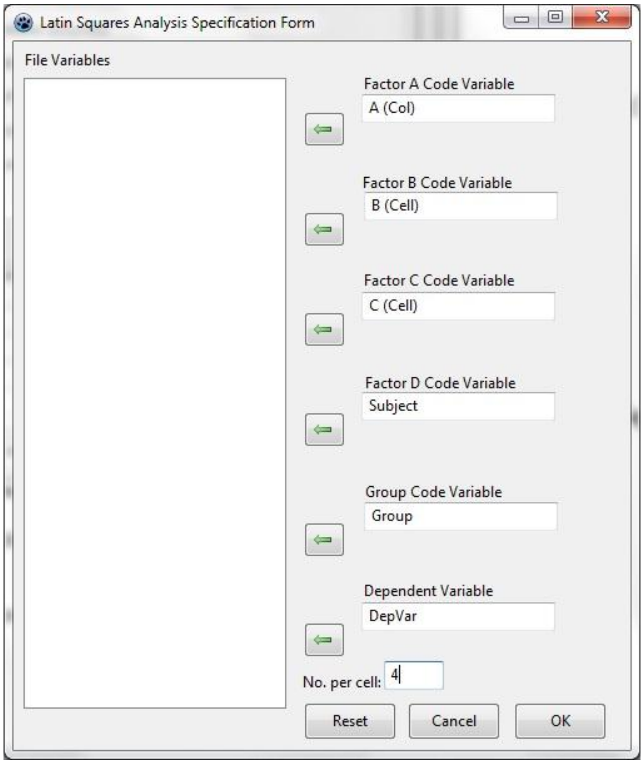


Fig. 6.18   Latin Squares Repeated Analysis Plan 7 (Superimposed Squares)

Latin Squares Repeated Analysis Plan 7 (superimposed squares)

Sums for ANOVA Analysis

Group (rows) times A Factor (columns) sums with 36 cases.

Variables

	1	2	3	Total
1	23.000	16.000	22.000	61.000
2	22.000	14.000	18.000	54.000
3	24.000	21.000	21.000	66.000
Total	69.000	51.000	61.000	181.000

Group (rows) times B (cells Factor) sums with 36 cases.

Variables

	1	2	3	Total
1	23.000	16.000	22.000	61.000
2	18.000	22.000	14.000	54.000
3	21.000	21.000	24.000	66.000
Total	62.000	59.000	60.000	181.000

Group (rows) times C (cells Factor) sums with 36 cases.

Variables

	1	2	3	Total
1	23.000	22.000	16.000	61.000
2	14.000	22.000	18.000	54.000
3	21.000	21.000	24.000	66.000
Total	58.000	65.000	58.000	181.000

Group (rows) times Subjects (columns) sums with 36 cases.

Variables

	1	2	3	4	Total
1	16.000	14.000	13.000	18.000	61.000
2	12.000	13.000	14.000	15.000	54.000
3	18.000	19.000	11.000	18.000	66.000
Total	46.000	46.000	38.000	51.000	181.000

Latin Squares Repeated Analysis Plan 7 (superimposed squares)

Source	SS	DF	MS	F	Prob.>F
Betw.Subj.	26.306	11			
Groups	6.056	2	3.028	1.346	0.308
Subj.w.g.	20.250	9	2.250		
Within Sub	70.667	24			
Factor A	13.556	2	6.778	2.259	0.133
Factor B	0.389	2	0.194	0.065	0.937
Factor C	2.722	2	1.361	0.454	0.642
residual	-	0	-		
Error w.	54.000	18	3.000		
Total	96.972	35			

Experimental Design for Latin Square

A (Col)	1	2	3
Group			
1	BC11	BC23	BC32
2	BC22	BC31	BC13
3	BC33	BC12	BC21

Cell means and totals

A (Col)	1	2	3	Total
Group				
1	5.750	4.000	5.500	5.083
2	5.500	3.500	4.500	4.500
3	6.000	5.250	5.250	5.500
Total	5.750	4.250	5.083	5.028

Means for each variable

A (Col)	1	2	3	Total
	5.750	4.250	5.083	5.028

B (Cell)	1	2	3	Total

```
               5.167      4.917      5.000      5.028
-----------------------------------------------------

-----------------------------------------------------
  C (Cell)     1          2          3          Total
-----------------------------------------------------
               4.833      5.417      4.833      5.028
-----------------------------------------------------

-----------------------------------------------------
     Group     1          2          3          Total
-----------------------------------------------------
               5.083      4.500      5.500      5.028
-----------------------------------------------------
```

Plan 9 Latin Squares

The sample data set labeled "LatinPlan9.LAZ" is used for the following analysis. The specification form shown below has the variables entered for the analysis. When you click the OK button, the results obtained are as shown following the form.

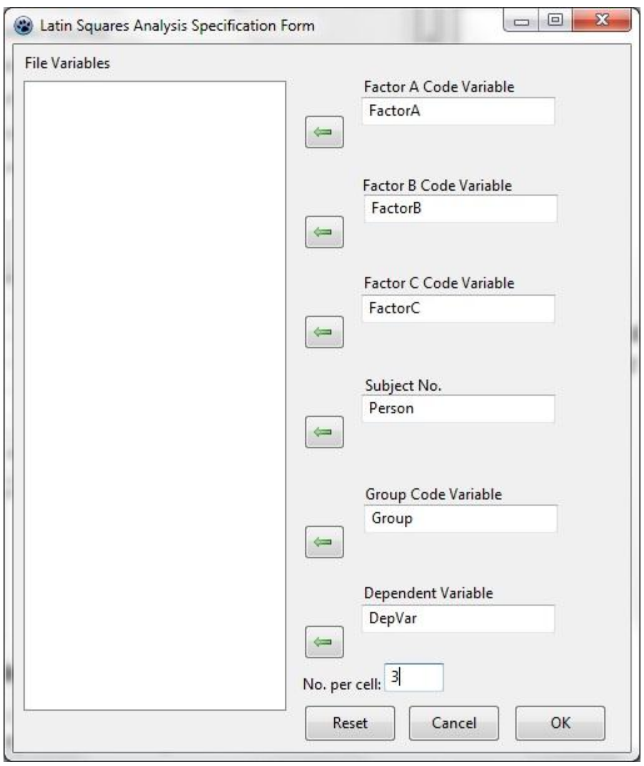


Fig. 6.19 Latin Squares Repeated Analysis Plan 9

```
Latin Squares Repeated Analysis Plan 9

Sums for ANOVA Analysis

ABC matrix

C level 1
                1          2          3
    1          13.000      3.000      9.000
    2           6.000      9.000      3.000
    3          10.000     14.000     15.000
```

C level 2

	1	2	3
1	18.000	14.000	18.000
2	19.000	24.000	20.000
3	8.000	11.000	10.000

C level 3

	1	2	3
1	17.000	12.000	20.000
2	14.000	13.000	9.000
3	15.000	12.000	17.000

AB sums with 27 cases.

Variables

	1	2	3	Total
1	48.000	29.000	47.000	124.000
2	39.000	46.000	32.000	117.000
3	33.000	37.000	42.000	112.000
Total	120.000	112.000	121.000	353.000

AC sums with 27 cases.

Variables

	1	2	3	Total
1	25.000	50.000	49.000	124.000
2	18.000	63.000	36.000	117.000
3	39.000	29.000	44.000	112.000
Total	82.000	142.000	129.000	353.000

BC sums with 27 cases.

Variables

	1	2	3	Total
1	29.000	45.000	46.000	120.000
2	26.000	49.000	37.000	112.000
3	27.000	48.000	46.000	121.000
Total	82.000	142.000	129.000	353.000

RC sums with 27 cases.

Variables

	1	2	3	Total
1	16.000	42.000	36.000	94.000
2	37.000	52.000	47.000	136.000
3	29.000	48.000	46.000	123.000
Total	82.000	142.000	129.000	353.000

Group totals with 27 valid cases.

Variables	1	2	3	4	5
	16.000	37.000	29.000	42.000	52.000

Variables	6	7	8	9	Total
	48.000	36.000	47.000	46.000	353.000

Subjects sums with 27 valid cases.

Variables	1	2	3	4	5
	7.000	9.000	14.000	28.000	15.000

Variables	6	7	8	9	10
	21.000	16.000	21.000	22.000	30.000

Variables	11	12	13	14	15
	28.000	19.000	10.000	19.000	23.000

Variables	16	17	18	Total
	25.000	28.000	18.000	0.000

Computation Terms
Term1 = 1538.383
term2 = 2811.000
term3 = 1541.074
term4 = 1540.185
term5 = 1612.185
term6 = 1581.889
term7 = 1712.556
term8 = 1619.667
term9 = 1769.667
term10 = 2575.000
term11 = 1651.000
term12 = 1572.630

Latin Squares Repeated Analysis Plan 9

Source	SS	DF	MS	F	Prob.>F
Betw.Subj.	1036.617	26			
Factor C	73.802	2	36.901	0.719	0.501
Rows	34.247	2	17.123	0.334	0.721
C x row	4.568	4	1.142	0.022	0.999
Subj.w.g.	924.000	18	51.333		
Within Sub	236.000	54			
Factor A	2.691	2	1.346	0.413	0.665

```
 Factor B     1.802         2     0.901     0.277     0.760
 Factor AC   97.679         4    24.420     7.492     0.000
 Factor BC    5.679         4     1.420     0.436     0.782
 AB prime     4.765         2     2.383     0.731     0.488
 ABC prime    6.049         4     1.512     0.464     0.762
 Error w.   117.333        36     3.259

Total      1272.617        80
```

```
Experimental Design for Latin Square

------------------------------
   FactorA   1    2    3
------------------------------
     Group
     1       B2   B3   B1
     2       B1   B2   B3
     3       B3   B1   B2
     4       B2   B3   B1
     5       B1   B2   B3
     6       B3   B1   B2
     7       B2   B3   B1
     8       B1   B2   B3
     9       B3   B1   B2
```

```
Latin Squares Repeated Analysis Plan 9

Means for ANOVA Analysis

ABC matrix

C level 1
            1          2          3
    1       4.333      1.000      3.000
    2       2.000      3.000      1.000
    3       3.333      4.667      5.000


C level 2
            1          2          3
    1       6.000      4.667      6.000
    2       6.333      8.000      6.667
    3       2.667      3.667      3.333


C level 3
            1          2          3
    1       5.667      4.000      6.667
    2       4.667      4.333      3.000
    3       5.000      4.000      5.667



AB Means with   81 cases.


Variables
                      1             2             3             4
         1       5.333         3.222         5.222         4.593
```

2	4.333	5.111	3.556	4.333
3	3.667	4.111	4.667	4.148
Total	4.444	4.148	4.481	4.358

AC Means with 81 cases.

Variables

	1	2	3	4
1	2.778	5.556	5.444	4.593
2	2.000	7.000	4.000	4.333
3	4.333	3.222	4.889	4.148
Total	3.037	5.259	4.778	4.358

BC Means with 81 cases.

Variables

	1	2	3	4
1	3.222	5.000	5.111	4.444
2	2.889	5.444	4.111	4.148
3	3.000	5.333	5.111	4.481
Total	3.037	5.259	4.778	4.358

RC Means with 81 cases.

Variables

	1	2	3	4
1	1.778	4.667	4.000	3.481
2	4.111	5.778	5.222	5.037
3	3.222	5.333	5.111	4.556
Total	3.037	5.259	4.778	4.358

Group Means with 81 valid cases.

Variables	1	2	3	4	5
	1.778	4.111	3.222	4.667	5.778

Variables	6	7	8	9	Total
	5.333	4.000	5.222	5.111	4.358

Subjects Means with 81 valid cases.

Variables	1	2	3	4	5
	2.333	3.000	4.667	9.333	5.000

```
Variables            6          7          8          9         10
                 7.000      5.333      7.000      7.333     10.000

Variables           11         12         13         14         15
                 9.333      6.333      3.333      6.333      7.667

Variables           16         17         18      Total
                 8.333      9.333      6.000      4.358
```

Analysis of Variance Using Multiple Regression Methods

A Comparison of ANOVA and Regression

In one-way analysis of variance with Fixed Effects, the model that describes the expected Y score is usually given as

$$Y_{i,j} = \mu + \alpha_j + e_{i,j} \tag{6.27}$$

where Yi,j is the observed dependent variable score for subject i in treatment group j,

μ is the population mean of the Y scores,

α_j is the effect of treatment j, and

$e_{i,j}$ is the deviation of subject i in the jth treatment group from the population mean for that group.

The above equation may be rewritten with sample estimates as

$$Y'_{i,j} = \overline{Y}_{..} + (\overline{Y}_{.j} - \overline{Y}_{..}) \tag{6.28}$$

For any given subject then, irrespective of group, we have

$$Y'_i = \overline{Y}_{..} + (\overline{Y}_{.1} - \overline{Y}_{..})X_1 + ... + (\overline{Y}_{.k} - \overline{Y}_{..})X_k \tag{6.29}$$

where X_j is 1 if the subject is in the group, otherwise 0.

If we let $B_0 = \overline{Y}_{..}$ and the effects $(\overline{Y}_{.j} - \overline{Y}_{..})$ be B_j for any group, we may rewrite the above equation as

$$Y'_i = B_0 + B_1X_1 + ... + B_kX_k \tag{6.30}$$

This is, of course, the general model for multiple regression! In other words, the model used in ANOVA may be directly translated to the multiple regression model. They are essentially the same model!

You will notice that in this model, each subject has K predictors X. Each predictor is coded a 1 if the subject is in the group, otherwise 0. If we create a variable for each group however, we do not have independence of the predictors. We lack independence because one group code is redundant information with the K-1 other group codes. For example, if there is only two groups and a subject is in group 1, then $X_1 = 1$ and X_2 MUST BE 0 since an individual cannot belong in both groups. There are only K-1 degrees of freedom for group membership - if an individual is not in groups 1 up to K we automatically know they belong to the Kth group. In order to use multiple regression, the predictor variables must be independent. For this reason, the number of predictors is restricted to one less than the number of groups. Since all α_j effects must sum to zero, we need only know the first K-1 effects - the last can be obtained by subtraction from 1 - $\Sigma\ \alpha_j$ where j = 1,..,K-1.

We also remember that

$$B_0 = \bar{Y}.. - (B_1\bar{X}_1 + ... + B_k\bar{X}_k). \tag{6.31}$$

Effect Coding

In order for B_0 to equal the grand mean of the Y scores, we must restrict our model in such a way that the sum of the products of the X means and regression coefficients equals zero. This may be done by use of "effect" coding. In this method there are K-1 independent variables for each subject. If a subject is in the group corresponding to the jth variable, he or she has a score $X_j = 1$ otherwise the score is $X_j = 0$. Subjects in the Kth group do not have a corresponding X variable so they receive a score of _1 in all of the group codes.

As an example, assume that you have 5 subjects in each of three groups. The "effect" coding of predictor variables would be

SUBJECT	Y	CODE 1	CODE 2	
01	5	1	0	
02	8	1	0	
03	4	1	0	(Group 1)
04	7	1	0	
05	3	1	0	
06	4	0	1	
07	6	0	1	
08	2	0	1	(Group 2)
09	9	0	1	
10	4	0	1	
11	3	-1	-1	
12	6	-1	-1	
13	5	-1	-1	(Group 3)
14	9	-1	-1	
15	4	-1	-1	

You may notice that the mean of X_1 and of X_2 are both zero. The cross-products of X_1X_2 is n_3, the size of the last group.

If we now perform a multiple regression analysis as well as a regular ANOVA for the data above, we will obtain the following results:

SOURCE	DF	SS	MS	F	PROB>F
Full Model	2	0.533	0.267	0.048	0.953
Groups	2	0.533	0.267	0.048	0.953
Residual	12	66.400	5.533		
Total	14	66.933			

R^2 = 0.008

You will note that the SS_{groups} may be obtained from either the ANOVA printout or the SS_{reg} in the Multiple Regression analysis. The SS_{error} is the same in both analyses as is the total sum of squares.

Orthogonal Coding

While effect coding provides the means of directly estimating the effect of membership in levels or treatment groups, the correlations among the independent variables are not zero, thus the inverse of that matrix may be difficult if done by hand. Of greater interest however, is the ability of other methods of data coding that permits the research to pre-specify contrasts or comparisons among particular treatment groups of interest. The method of orthogonal coding has several benefits:

I. The user can pre-plan comparisons among selected groups or treatments, and

II. the inter-correlation matrix is a diagonal matrix, that is, all off-diagonal values are zero. This results in a solution for the regression coefficients which can easily be calculated by hand.

When orthogonal coding is utilized, there are K-1possible orthogonal comparisons in each factor. For example, if there are four treatment levels of Factor A, there are 3 possible orthogonal comparisons that may be made among the treatment means. To illustrate orthogonal coding, we will utilize the same example as before. The previous effect coding will be replaced by orthogonal coding as illustrated in the data below:

```
SUBJECT   Y        CODE 1        CODE 2

01        5        1             1
02        8        1             1
03        4        1             1            (Group 1)
04        7        1             1
05        3        1             1

06        4        -1            1
07        6        -1            1
08        2        -1            1            (Group 2)
09        9        -1            1
10        4        -1            1

11        3        0             -2
12        6        0             -2
13        5        0             -2           (Group 3)
14        9        0             -2
15        4        0             -2
```

Now notice that, as before, the sum of the values in each coding vector is zero. Also note that, in this case, the product of the coding vectors is also zero. (Multiply the code values of two vectors for each subject and add up the products - they should sum to zero.) Vector 1 above (Code 1) represents a comparison of treatment group 1 with treatment group 2. Vector 2 represents a comparison of groups 1 AND 2 with group 3.

Now let us look at coding for, say, 5 treatment groups. The coding vectors below might be used to obtain orthogonal contrasts:

```
GROUP VECTOR 1       VECTOR 2        VECTOR 3        VECTOR 4
  1       1              1               1               1

  2      -1              1               1               1

  3       0             -2               1               1

  4       0              0              -3               1
```

5	0	0	0	-4

As before, the sum of coefficients in each vector is zero and the product of any two vectors is also zero. This assumes that there are the same number of subjects in each group. If groups are different in size, one may use additional multipliers based on the proportion of the total sample found in each group. The treatment group number in the left column may, of course, represent any one of the treatment groups thus it is possible to select a specific comparison of interest by assigning the treatment groups in the order necessary to obtain the comparison of interest.

Return now to the previous example. The results from the regression analysis program as well as the ANOVA program are presented in the Fig.s below. The first Fig. presents the inter-correlation matrix among the variables. Notice that the inter-correlations among the coding vectors are zero. The next Fig. presents the R^2 and the summary of regression coefficients. Multiplication of the R^2 times the sum of squares for the dependent variable will yield the sum of squares for regression. This will equal the sum of squares for groups in the subsequent ANOVA results table. By use of orthogonal vectors, we may also note that the regression coefficients are simply the correlation of each vector with the dependent variable. Multiplication of the squared regression coefficients times the sum of squares total will therefore give the sum of squares due to each contrast. The total sum of squares for groups is simply the sum of the sum of squares for each contrast! The test of departure of the regression coefficients from zero is a test of significance for the contrast in the corresponding coding vector. The a priori specified contrasts, unlike post-hoc comparisons maintain the selected alpha rate and more power. Hence, sensitivity to true population treatment effects are more likely to be detected by the planned comparison than by a post-hoc comparison.

Dummy Coding

Effect and orthogonal coding methods both resulted in code vectors which summed to zero across the subjects. In each of those cases, the constant B_0 estimates the population mean since it is the grand mean of the sample (see equation 9). Both methods of coding also resulted in the same squared multiple correlation coefficient R^2 indicating that the proportion of variance explained by both methods is the same.

Another method of coding which is popular is called "dummy" coding. In this method, K-1 vectors are also created for the coding of membership in the K treatment groups. However, the sum of the coded vectors do not add to zero as in the previous two methods. In this coding scheme, if a subject is a member of treatment group 1, the subject receives a code of 1. All other treatment group subjects receive a code of 0. For a second vector (where there are more than two treatment groups), subjects that are in the second treatment group are coded with a 1 and all other treatment group subjects are coded 0. This method continues for the K-1 groups. Clearly, members of the last treatment group will have a code of zero in all vectors. The coding of members in each of five treatment groups is illustrated below:

GROUP	VECTOR 1	VECTOR 2	VECTOR 3	VECTOR 4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	0	0	0

With this method of coding, like that of effect coding, there will be correlations among the coding vectors which differ from zero thus necessitating the computation of the inverse of a symmetric matrix rather than a diagonal matrix. Never the less, the squared multiple correlation coefficient R^2 will be the same as with the other coding methods and therefore the SS_{reg} will again reflect the treatment effects. Unfortunately, the resulting regression coefficients reflect neither the direct effect of each treatment or a comparison among treatment groups. In addition, the constant B_0 reflects the mean only of the treatment group (last group) which receives all zeroes in the coding vectors. If however, the overall effects of treatment is the finding of interest, dummy coding will give the same results.

Two Factor ANOVA by Multiple Regression

In the above examples of effect, orthogonal and dummy coding of treatments, we dealt only with levels of a single treatment factor. We may, however, also analyze multiple factor designs by multiple regression using each of these same coding methods. For example, a two-way analysis of variance using two treatment factors will typically provide the test of effects for the A factor, the B factor and the interaction of the A and B treatments. We will demonstrate the use of effect, orthogonal and dummy coding for a typical research design involving three levels of an A treatment and four levels of a B treatment.

Example Design

Levels of Treatment B

		1	2	3	4
Levels	1				
of	2				
Treatment	3				
A					

For effect coding in the above design, we apply effect codes to the A treatment levels first and then, beginning again, to the B treatment levels independently of the A codes. Finally, we multiple each of the code vectors of the A treatments times each of the code vectors of the B treatment to create the interaction vectors. The vectors below illustrate this for the above design:

		A		B			A x B					
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
ROW	COL	A_1	A_2	B_1	B_2	B_3	A_1B_1	A_1B_2	A_1B_3	A_2B_1	A_2B_2	A_2B_3
1	1	1	0	1	0	0	1	0	0	0	0	0
1	2	1	0	0	1	0	0	1	0	0	0	0
1	3	1	0	0	0	1	0	0	1	0	0	0
1	4	1	0	-1	-1	-1	-1	-1	-1	0	0	0
2	1	0	1	1	0	0	0	0	0	1	0	0
2	2	0	1	0	1	0	0	0	0	0	1	0
2	3	0	1	0	0	1	0	0	0	0	0	1
2	4	0	1	-1	-1	-1	0	0	0	-1	-1	-1

3	1	-1	-1	1	0	0	-1	0	0	-1	0	0
3	2	-1	-1	0	1	0	0	-1	0	0	-1	0
3	3	-1	-1	0	0	1	0	0	-1	0	0	-1
3	4	-1	-1	-1	-1	-1	1	1	1	1	1	1

If you add the values in any one of the vectors above you will see they sum to zero. In addition, the product of any two vectors selected from a combination of treatment A, B or AxB sets will also be zero! With effect coding, the treatment effect vectors from one factor are orthogonal (uncorrelated) with the treatment effect vectors of the other factor as well as the interaction effect vectors. The effect vectors within each treatment or interaction are not, however, orthogonal.

With effect coding, we may "decompose" the R^2 for the full model into the three separate parts, that is

$$R^2_{y.1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11} = R^2_{y.1\,2} + R^2_{y.3\,4\,5} + R^2_{y.6\,7\,8\,9\,10\,11} \qquad (6.32)$$

since the A, B and AxB effects are orthogonal.

Again, the regression coefficients directly report the effect of treatment group membership, that is, B_1 is the effect of treatment group 1 in the A factor and B_2 is the effect of treatment group 2 in the A factor. The effect of treatment group 3 in the A factor can be obtained as

$$\alpha_3 = 1 - \Sigma(\alpha_1 + \alpha_2) = 1 - (B_1 + B_2) \qquad (6.33)$$

since the sum of effects is constrained to equal zero. Similarly, B_3 estimates β_1, B_4 estimates β_2 and B_5 estimates the B factor effect β_3 for column 3. The effect of column four is also obtained as before, that is,

$$\beta_4 = 1 - (B_3 + B_4 + B_5). \qquad (6.34)$$

The interaction effects for the cells, $\alpha\beta_{ij}$, may be obtained from the regression coefficients corresponding to the interaction vectors. In this example, B_6 estimates $\alpha\beta_{11}$, B_7 estimates $\alpha\beta_{12}$, B_8 estimates $\alpha\beta_{13}$, B_9 estimates $\alpha\beta_{21}$, B_{10} estimates $\alpha\beta_{22}$ and B_{11} estimates $\alpha\beta_{23}$. Since the sum of the interaction effects in any row or column must be zero, we can determine estimates for the cells in rows 1 and 2 of column 4 as follows:

$$\alpha\beta_{14} = 1 - (B_6 + B_7 + B_8) \text{ and} \qquad (6.35)$$

$$\alpha\beta_{24} = 1 - (B_9 + B_{10} + B_{11}). \qquad (6.36)$$

We may also utilize orthogonal coding vectors within each treatment factor as we did for effect coding above. The same two-factor design above could utilize the vectors below:

		A		B			A x B					
		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
ROW	COL	A1	A2	B1	B2	B3	A1B1	A1B2	A1B3	A2B1	A2B2	A2B3
1	1	1	1	1	1	1	1	1	1	1	1	1
1	2	1	1	-1	1	1	-1	1	1	-1	1	1
1	3	1	1	0	-2	1	0	-2	1	0	-2	1

1	4	1	1	0	0	-3	0	0	-3	0	0	-3
2	1	-1	1	1	1	1	-1	-1	-1	1	1	1
2	2	-1	1	-1	1	1	1	-1	-1	-1	1	1
2	3	-1	1	0	-2	1	0	2	-1	0	-2	1
2	4	-1	1	0	0	-3	0	0	3	0	0	-3
3	1	0	-2	1	1	1	0	0	0	-2	-2	-2
3	2	0	-2	-1	1	1	0	0	0	2	-2	-2
3	3	0	-2	0	-2	1	0	0	0	0	4	-2
3	4	0	-2	0	0	-3	0	0	0	0	0	6

As before, the sum of each vector is zero. This time however, the product of vectors within each factor as well as between factors and interaction are zero. All vectors are orthogonal to one another. The inter-correlation matrix is therefore a diagonal matrix and easily inverted by hand. The R^2 for the full model may be easily decomposed into the sum of squared simple correlations between the dependent and independent score vectors, that is

$$R^2_{y.1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11} =$$

$$r^2_{y.1} + r^2_{y.2} + \qquad \text{(row effects)}$$

$$r^2_{y.3} + r^2_{y.4} + r^2_{y.5} + \qquad \text{(column effects)}$$

$$r^2_{y.6} + r^2_{y.7} + r^2_{y.8} + r^2_{y.9} + r^2_{y.10} + r^2_{y.11} \qquad \text{(interaction effects)} \qquad (6.37)$$

The regression coefficients obtained with orthogonal coding vectors represent planned comparisons among treatment means. Using the coding vectors for this example, the B_1 coefficient would represent the comparison of row 1 mean with row 2 mean. B_2 would represent the contrast of row 3 mean with the combination of rows 1 and 2. The coefficients B_3, B_4 and B_5 similarly contrast column means. The contrasts represented by the interaction vectors will reflect comparisons among specific cell combinations. For example, B_7 above will reflect a contrast of the combined cells in row 1 column 1 and row 2 column 2 with the combined cells of row 1 column 2 and row 2 column 1.

Analysis of Covariance By Multiple Regression Analysis

In the previous sections we have examined methods for coding nominal variables of analysis of variance designs to explain the variance of the continuous dependent variable. We may, however, also include one or more independent variables that are continuous and expected to have the same correlation with the dependent variable in each treatment group population. As an example, assume that the two-way ANOVA design discussed in the previous section represents an experiment in which Factor A represent three type of learning reinforcement (positive only, negative only and combined positive and negative) while Factor B represents four types of learning situations (CAI, teacher led, self instruction, and peer tutor). Assume the dependent variable is a standardized measure of Achievement in learning the French language. Finally, assume the treatment groups are exposed to the treatments for a sufficiently long period of time to produce measurable achievement by most students and that the students have been randomly assigned to the treatment groups. It may occur to the reader that achievement in learning a new language might be related to general intelligence as measured, say, by the Stanford-Binet Intelligence Test as well as related to prior English achievement measured by a standardized achievement test in English. Variation in IQ and English achievement of subjects in the treatment groups may explain a portion of the within treatment cell variance. We prefer to have the within cell variance as small as possible since it is the basis of the mean squared residual used in the F tests of our treatment effects. To accomplish this, we can first extract that portion of total dependent score variance explained by IQ and English achievement before examining that portion of the remaining variance explainable by our main treatment effects. Assume therefore, that in addition to the eleven vectors representing Factor A level effects, Factor B level effects and Factor interaction effects, we include X_{12} and X_{13} predictors of IQ and English. Then the proportion of variance for Factor A effects controlling for IQ and English is

$$R^2_{y\cdot 1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11\,12\,13} - R^2_{y\cdot 3\,4\,5\,6\,7\,8\,9\,10\,11\,12\,13}$$

The proportion of French achievement variance due to Factor B treatments controlling for IQ and English would be

$$R^2_{y.1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11\,12\,13} - R^2_{y.1\,2\,6\,7\,8\,9\,10\,11\,12\,13}$$

and the proportion of variance due to interaction of Factor A and Factor B controlling for IQ and English would be

$$R^2_{y.1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11\,12\,13} - R^2_{y.1\,2\,3\,4\,5\,12\,13}$$

In each of the above, the full model contains all predictors while the restricted model contains all variables except those of the effects being evaluated. The F statistic for testing the hypothesis of equal treatment effects is

$$F = \frac{R^2_{full} - R^2_{restricted}}{1.0 - R^2_{full}} \cdot \frac{N - K_f - 1}{K_f - K_r} \tag{6.38}$$

where K_f is the number of predictors in the full model, and

K_r is the number of predictors in the restricted model.

The numerator and denominator degrees of freedom for these F statistics is $(K_f - K_r)$ and $(N - K_f - 1)$ respectively.

Analysis of Covariance assumes homogeneity of covariance among the treatment groups (cells) in the populations from which the samples are drawn. If this assumption holds, the interaction of the covariates with the main treatment factors (A and B in our example) should not account for significant variance of the dependent variable. You can explicitly test this assumption therefore by constructing a full model which has all of the previously included independent variables plus prediction vectors obtained by multiplying each of the treatment level vectors times each of the covariates. In our above example, for instance, we would multiply each of the first five vectors times both IQ and English vectors (X_{12} and X_{13}) resulting in a full model with 10 more variables (23 predictors in all).

The R^2 from our previous full model would be subtracted from the R^2 for this new full model to determine the proportion of variance attributable to heteroscedasticity of the covariance among the treatment groups. If the F statistic for this proportion is significant, we cannot employ the analysis of covarance model. The implication would be that somehow, IQ and prior English achievement interacts differently among the levels of the treatments. Note that in testing this assumption of homogeneity of covariance, we have a fairly large number of variables in the regression analysis. To obtain much power in our F test, we need a considerable number of subjects. Several hundred subjects would not be unreasonable for this study, i.e. 25 subjects per each of the eight treatment groups!

Sums of Squares by Regression

The General Linear Model

We have seen in the above discussion that the multiple regression method may be used to complete an analysis of variance for a single dependent variable. The model for multiple regression is:

$$y_i = \sum_{j=1}^{k} B_j X_j + e_i \tag{6.39}$$

where the jth B value is a coefficient multiplied times the jth independent predictor score, Y is the observed dependent score and e is the error (difference between the observed and the value predicted for Y using the sum of weighted independent scores.)

In some research it is desirable to determine the relationship between multiple dependent variables and multiple independent variables. Of course, one could complete a multiple regression analysis for each dependent variable but this would ignore the possible relationships among the dependent variables themselves. For example, a teacher might be interested in the relationship between the sub-scores on a standardized achievement test (independent variables) and the final examination results for several different courses (dependent variables.) Each of the final examination scores could be predicted by the sub-scores in separate analyses but most likely the interest is in knowing how well the sub-scores account for the combined variance of the achievement scores. By assigning weights to each of the dependent variables as well as the independent variables in such a way that the composite dependent score is maximally related to the composite independent score we can quantify the relationship between the two composite scores. We note that the squared product-moment correlation coefficient reflects the proportion of variance of a dependent variable predicted by the independent variable.

We can express the model for the general linear model as:

$$YM = BX + E \tag{6.40}$$

where Y is an n (the number of subjects) by m (the number of dependent variables) matrix of dependent variable values, M is a m by s (number of coefficient sets), X is a n by k (the number of independent variables) matrix, B is a k by s matrix of coefficients and E is a vector of errors for the n subjects.

The General Linear Model (GLM) procedure is an analysis procedure that encompasses a variety of analyses. It may incorporate multiple linear regression as well as canonical correlation analysis as methods for analyzing the user's data. In some commercial statistics packages the GLM method also incorporates non-linear analyses, maximum-likelihood procedures and a variety of tests not found in the current version of this model. The version in LazStats is currently limited to a single dependent variable (continuous measure.) You should complete analyses with multiple dependent variables with the Canonical Correlation procedure.

One can complete a variety of analyses of variance with the GLM procedure including multiple factor ANOVA and repeated and mixed model ANOVAs.

The output of the GLM can be somewhat voluminous in that the effects of treatment variables and covariates are analyzed individually by comparing regression models with and without those variables.

Analysis of Variance Using Multiple Regression Methods

An Example of an Analysis of Covariance

We will demonstrate the analysis of covariance procedure using multiple regression by loading the file labeled “Ancova2.LAZ”. In this file we have a treatment group code for four groups, a dependent variable (X) and two covariates (Y and Z.) The procedure is started by selection the “Analysis of Covariance by Regression” option in the Comparisons sub-menu under the Statistics menu. Shown below is the completed specification form for our analysis:

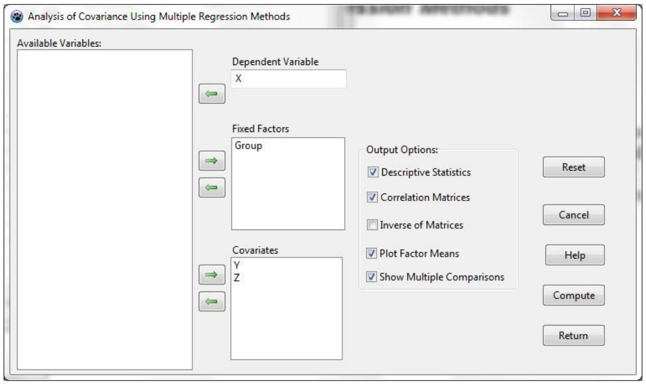


Fig. 6.20   Analysis of Covariance Form

When we click the Compute button, the following results are obtained:

```
ANALYSIS OF COVARIANCE USING MULTIPLE REGRESSION

File Analyzed: C:\Users\wgmiller\LazStats\LazStatsData\ANCOVA2.LAZ


Model for Testing Assumption of Zero Interactions with Covariates


Correlation Matrix with   40 cases.
```

Variables

	Y	Z	A1	A2	A3
Y	1.000	0.547	-0.199	0.062	0.212
Z	0.547	1.000	-0.154	-0.048	-0.077
A1	-0.199	-0.154	1.000	0.500	0.500
A2	0.062	-0.048	0.500	1.000	0.500
A3	0.212	-0.077	0.500	0.500	1.000
YxA1	-0.196	-0.157	0.989	0.519	0.519
YxA2	0.079	-0.045	0.487	0.988	0.487
YxA3	0.221	-0.080	0.472	0.472	0.990
ZxA1	-0.188	-0.210	0.968	0.510	0.510
ZxA2	0.061	-0.107	0.493	0.970	0.493
ZxA3	0.190	-0.102	0.495	0.495	0.964
X	0.697	0.653	0.088	0.018	0.053

Variables

	YxA1	YxA2	YxA3	ZxA1	ZxA2
Y	-0.196	0.079	0.221	-0.188	0.061
Z	-0.157	-0.045	-0.080	-0.210	-0.107
A1	0.989	0.487	0.472	0.968	0.493
A2	0.519	0.988	0.472	0.510	0.970
A3	0.519	0.487	0.990	0.510	0.493
YxA1	1.000	0.516	0.501	0.980	0.522
YxA2	0.516	1.000	0.469	0.508	0.980
YxA3	0.501	0.469	1.000	0.493	0.476
ZxA1	0.980	0.508	0.493	1.000	0.543
ZxA2	0.522	0.980	0.476	0.543	1.000
ZxA3	0.524	0.493	0.972	0.545	0.527
X	0.069	0.029	0.056	0.069	0.020

Variables

	ZxA3	X
Y	0.190	0.697
Z	-0.102	0.653
A1	0.495	0.088
A2	0.495	0.018
A3	0.964	0.053
YxA1	0.524	0.069
YxA2	0.493	0.029

YxA3	0.972	0.056
ZxA1	0.545	0.069
ZxA2	0.527	0.020
ZxA3	1.000	0.038
X	0.038	1.000

MEANS with 40 valid cases.

Variables	Y	Z	A1	A2	A3
	17.550	14.675	0.000	0.000	0.000

Variables	YxA1	YxA2	YxA3	ZxA1	ZxA2
	-0.400	0.125	0.425	-0.400	-0.125

Variables	ZxA3	X
	-0.200	7.125

VARIANCES with 40 valid cases.

Variables	Y	Z	A1	A2	A3
	8.254	13.866	0.513	0.513	0.513

Variables	YxA1	YxA2	YxA3	ZxA1	ZxA2
	144.349	163.599	174.302	116.759	125.035

Variables	ZxA3	X
	124.113	4.163

STD. DEV.S with 40 valid cases.

Variables	Y	Z	A1	A2	A3
	2.873	3.724	0.716	0.716	0.716

Variables	YxA1	YxA2	YxA3	ZxA1	ZxA2
	12.015	12.791	13.202	10.806	11.182

Variables	ZxA3	X
	11.141	2.040

Analysis of Variance for the Model to Test Regression Homogeneity

SOURCE	Deg.F.	SS	MS	F	Prob>F
Explained	11	121.43	11.04	7.550	0.0000
Error	28	40.94	1.46		
Total	39	162.38			

R Squared = 0.748

Model for Analysis of Covariance

Correlation Matrix with 40 cases.

Variables

	Y	Z	A1	A2	A3
Y	1.000	0.547	-0.199	0.062	0.212
Z	0.547	1.000	-0.154	-0.048	-0.077
A1	-0.199	-0.154	1.000	0.500	0.500
A2	0.062	-0.048	0.500	1.000	0.500
A3	0.212	-0.077	0.500	0.500	1.000
X	0.697	0.653	0.088	0.018	0.053

Variables

	X
Y	0.697
Z	0.653

```
        A1      0.088
        A2      0.018
        A3      0.053
         X      1.000



MEANS with   40 valid cases.

Variables            Y            Z           A1           A2           A3
                17.550       14.675        0.000        0.000        0.000

Variables            X
                 7.125


VARIANCES with   40 valid cases.

Variables            Y            Z           A1           A2           A3
                 8.254       13.866        0.513        0.513        0.513

Variables            X
                 4.163


STD. DEV.S with   40 valid cases.

Variables            Y            Z           A1           A2           A3
                 2.873        3.724        0.716        0.716        0.716

Variables            X
                 2.040


Test for Homogeneity of Group Regression Coefficients
Change in R2 = 0.0462. F =      0.854  Prob.> F = 0.5398 with d.f.        6 and       28

R Squared =        0.702

Analysis of Variance for the ANCOVA Model
    SOURCE      Deg.F.      SS          MS          F           Prob>F
 Explained           5      113.94       22.79      15.996      0.0000
     Error          34       48.44        1.42
     Total          39      162.38


Unadjusted Group Means for Group Variables Group

Means with   40 valid cases.

Variables
                 7.400        7.000        7.200        6.900


Intercepts for Each Group Regression Equation for Variable: Group

Inercepts with   40 valid cases.

Variables      Group 1      Group 2      Group 3      Group 4
                -2.489       -4.065       -4.363       -4.028


Adjusted Group Means for Group Variables Group

Means with   40 valid cases.

Variables      Group 1      Group 2      Group 3      Group 4
                 8.373        6.796        6.498        6.833
Multiple Comparisons Among Group Means
```

```
Comparison of Group   1 with Group   2
F =      8.017, probability = 0.008 with degrees of freedom     1 and    34
Comparison of Group   1 with Group   3
F =      9.834, probability = 0.004 with degrees of freedom     1 and    34
Comparison of Group   1 with Group   4
F =     16.025, probability = 0.000 with degrees of freedom     1 and    34
Comparison of Group   2 with Group   3
F =      0.297, probability = 0.590 with degrees of freedom     1 and    34
Comparison of Group   2 with Group   4
F =      1.296, probability = 0.263 with degrees of freedom     1 and    34
Comparison of Group   3 with Group   4
F =      0.310, probability = 0.581 with degrees of freedom     1 and    34

Test for Each Source of Variance - Type III SS
----------------------------------------------------------------------
    SOURCE      Deg.F.      SS          MS          F           Prob>F
----------------------------------------------------------------------

      Cov0           1       38.11       38.11      26.754       0.0000

      Cov1           1       12.50       12.50       8.778       0.0055

         A           3       17.95        5.98       4.200       0.0124
----------------------------------------------------------------------
     ERROR          34       48.44        1.42
----------------------------------------------------------------------
```

The results reported above begin with a regression model that includes group coding for the four groups (A1, A2 and A3) and again note that the fourth group is automatically identified by members NOT being in one of the first three groups. This model also contains the covariates X and Z as well as the cross-products of group membership and covariates. By comparing this model with the second model created (one which leaves out the cross-products of groups and covariates) we can assess the degree to which the assumptions of homogeneity of covariance among the groups is met. In this particular example, the change in the R2 from the full model to the restricted model was quite small and we conclude that the assumption of homogeneity of covariance is reasonable. The analysis of variance model for the restricted model indicates that the X covariate is probably contributing significantly to the explained variance of the dependent variable Y. The tests for each source of variance at the end of the report confirms that not only are the covariates related to Y but that the group effects are also significant. The comparisons of the group means following adjustment for the covariate effects indicate that group 1 differs from groups 2 and 3 and that group 3 appears to differ from group 4.

Chapter 7. Multivariate Statistics

Canonical Correlation

Introduction

Canonical correlation analysis involves obtaining an index that describes the degree of relationship between two variables, each of which is a weighted composite of other variables. We have already examined the situation of an index between one variable and a weighted composite variable when we studied the multiple correlation coefficient of chapter X. Using a form similar to that used in multiple regression analysis, we might consider:

$$\beta_{y1}Y_1 + \beta_{y2}Y_2 + .. + \beta_{ym}Y_m + \beta_y = \beta_{x1}X_1 + .. + \beta_{xn}X_n + \beta_x$$

as a model for the regression of the composite function Y_c on the composite function Xc where

$$Y_c = \sum_{i=1}^{m} \beta_{yi}Y_i \quad \text{and } X_c = \sum_{j=1}^{n} \beta_{yj}X_j \qquad (7.1)$$

and the Y and X scores are in standardized form (z scores).

Using 'least-squares' criteria, we can maximize the simple product-moment correlation between Yc and Xc by selecting coefficients (Betas) which minimize the residuals (e). As in multiple regression, this involves solving partial derivatives for the β's on each side of the equation. The least-squares solution is more complicated than for multiple regression and will not be covered in this text. (See T.W. Anderson, An Introduction to Multivariate Analysis, 1958, chapter 12.)

Unfortunately for the beginning student, the canonical correlation analysis does not yield just one correlation index (R_c), but in fact may yield up to m or n (whichever is smaller) independent coefficients. This is because there are additional linear functions of the X's and Y's which may "explain" the residual variances $\square_y$ and $\square_x$ not explained by the first set of βx and βy weights. Each set of these canonical functions explains an additional portion of the common variance of the X and Y variables!

Before introducing the mathematics of obtaining these canonical correlations, the sets of corresponding weights and statistical tests of significance, we need to have a basic understanding of the concept of roots and vectors of a matrix.

Eigenvalues and Eigenvectors

A concept which occurs frequently in multivariate statistical analyses is the concept of eigenvalues (roots) and associated eigenvectors. Canonical correlation, factor analysis, multivariate analysis of variance, discriminant analysis, etc. utilize the roots and vectors of matrices in their solutions. To understand this concept, consider a k by k matrix (e.g. a correlation matrix)$[R]_{kxk}$. A basic problem in mathematical statistics is to find a k x 1 vector (matrix) $[E]_j$ and a scalar (single value) y_j such that

$$\underset{kxk}{[R]} \; \underset{kx1}{[E]} = \underset{j}{y} \, \underset{kx1}{[E]} \quad \text{where at least one element} \qquad (7.2)$$

of $\underset{kx1}{[E]}$ is not zero.

This equation may be rewritten as

$$[R]_{kxk} [E]_{kx1} - y_j [E]_{kx1} = [0]_{kx1}$$

or as $$([R]_{kxk} - y_j [I]_{kxk}) [E]_{kx1} = [0]_{kx1} \quad (7.3)$$

If we were to solve this equation for [E] by multiplying both sides of the last equation by the inverse of the matrix in the parenthesis (assuming the inverse exists), then [E] would be zero, a solution which violates our desire that at least one element of [E] NOT be zero! Consequently, [E] will have a non-zero element only if the determinant of

$$([R]_{kxk} - y_j [I]_{kxk})$$

is zero. The equation

$$|[R]_{kxk} - y_j [I]_{kxk}| = 0 \quad (7.4)$$

is called the characteristic equation. The properties of this equation have many applications in science and engineering.

The vector $[E]_{kx1}$ and the scalar y_j in the equation (5.43) are the eigenvector and eigenvalue of the matrix $[R]_{kxk}$.

Eigenvalues and eigenvectors are also known as characteristic roots and vectors of a matrix. To demonstrate that the eigenvalue is a root of a characteristic equation, consider the simple case of a 2x2 matrix such as

$$\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}$$

The problem is to find the root y_j in solving

$$\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} \cdot \begin{vmatrix} e_1 \\ e_2 \end{vmatrix} = y_j \begin{vmatrix} e_1 \\ e_2 \end{vmatrix}$$

Using the determinant:

$$\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} - \begin{vmatrix} y & 0 \\ 0 & y \end{vmatrix} = 0$$

or

$$\left| \begin{vmatrix} b_{11}\_y & b_{12} \\ b_{21} & b_{22}\_y \end{vmatrix} \right| = 0$$

This determinant has the solution

$$(b_{11} - y)(b_{22} - y) - b_{12} b_{21} = 0$$

or $b_{11} b_{22} - yb_{22} - yb_{11} + y^2 - b_{12} b_{21} = 0$

or $y^2 - y(b_{22} + b_{11}) + (b_{11} b_{22} - b_{12} b_{21}) = 0$

This is a quadratic equation with two roots y_1 and y_2 given by

$$.5 \{(b_{22} + b_{11}) +/- [(b_{22} + b_{11})^2 - 4(b_{11} b_{22} - b_{12} b_{21}).5 \}]$$

With the roots y_1 and y_2 evaluated, the elements e_1 and e_2 of the eigenvector can be solved from

$$\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} \begin{vmatrix} e_1 \\ e_2 \end{vmatrix} = y_j \begin{vmatrix} e_1 \\ e_2 \end{vmatrix}$$

which reduces to the equations (for each root):

$$b_{11} e_1 + b_{12} e_2 = y e_1$$

$$b_{21} e_1 + b_{22} e_2 = y e_2$$

and further reduces to

$$(b_{11} - y)e_1 + b_{12} e_2 = 0$$

$$b_{21}e_1 + (b_{22} - y)e_2 = 0$$

Solving these last equations simultaneously for e_1 and e_2 will yield the elements of the eigenvector [E].

There will be an eigenvector for each eigenvalue. In the case of the 2x2 matrix, the complete solution will be

$$\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} \begin{vmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{vmatrix} = \begin{vmatrix} y_1 & 0 \\ 0 & y_2 \end{vmatrix} \begin{vmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{vmatrix} \quad (7.5)$$

Every kxk matrix will have as many eigenvalues and eigenvectors as its order. Not all of the eigenvalues may be nonzero. When a square matrix [R] is symmetric, its eigenvalues are all real and the associated eigenvectors are orthogonal (products equal zero). If some of the eigenvalues are zero, we say that the RANK of the matrix is (k - p) where p is the number of roots equal to zero. The TRACE of a symmetric matrix is the sum of the eigenvalues. The determinant of the matrix is the product of all roots.

Other relationships obtainable from symmetric matrices are:

$$\underset{kxk}{[R]} \; \underset{kxk}{[E]} = \underset{kxk}{[y]} \; \underset{kxk}{[E]} \quad (7.6)$$

$$c\underset{kxk}{[R]} \; \underset{kxk}{[E]} = c\underset{kxk}{[y]} \; \underset{kxk}{[E]} \quad \text{where c is a constant.}$$

It may be pointed out that for any symmetric matrix and its eigenvalues there may be an infinite number of associated eigenvector matrices. There is, however, at least one matrix of eigenvectors that is *orthonormal*. An orthonormal matrix is one which when premultiplied by its transpose yields an identity matrix. If [E] is orthonormal then:

$$\underset{k\times k}{[E]'}\ \underset{k\times k}{[E]} = \underset{k\times k}{[I]} \tag{7.7}$$

and $$\underset{k\times k}{[E]'} = \underset{k\times k}{[E]^{-1}} \tag{7.8}$$

The Canonical Analysis

In completing a canonical analysis, the inter-correlation matrix among all of the variables may be partitioned into four sub-matrices as shown symbolically below. The $[R_{11}]$ matrix is the matrix of correlations among the "left_hand" variables of the equation presented earlier. The $[R_{22}]$ matrix is the correlations among the "right_hand" variables of our model. $[R_{12}]$ are the inter-correlations among the left and right hand variables. $[R_{21}]$ is the transpose of $[R_{12}]$.

$$[R] = \begin{vmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{vmatrix} \tag{7.9}$$

To start the canonical analysis, a product matrix is first formed by:

$$[R_p] = [R_{22}]^{-1} [R_{21}] [R_{11}]^{-1} [R_{12}] \tag{7.10}$$

The equation

$$([R_p] - y_j[I])v_j = 0 \tag{7.11}$$

where y_j is the jth root and v_j is the corresponding eigenvector is solved using the characteristic equation:

$$| [R_p] - y_j[I] | = 0 \tag{7.12}$$

with the restriction that

$$[V]'[R_{22}][V] = [I] \tag{7.13}$$

The canonical correlation ${}_1R_c$ corresponding to the first linear relationship between the left hand variables and the right hand variables is equal to the square root of the first root y_1. In general, the jth canonical correlation is obtained as:

$${}_jR_c = \sqrt{y_j} \tag{7.14}$$

The canonical correlation may be interpreted as the product-moment correlation between a weighted composite of the left-hand variables and a weighted composite of the right-hand variables.

Discriminant Function / MANOVA

Theory

Multiple discriminant function analysis is utilized to obtain a set of linear functions which maximally discriminate (differentiate) among subjects belonging to several different groups or classifications. For example, an investigator may want to develop equations which differentiate among successful occupational groups based on responses to items of a questionnaire. The functions obtained may be written as:

$$F_j = B_{j,1}X_1 + ... + B_{j,m}X_m \tag{7.15}$$

where

X_i represents an observed variable (i= 1..m),
$B_{j,i}$ is a coefficient for the X_i variable from the
j<u>th</u> discriminant function

The coefficients of these discriminant functions are the normalized vectors corresponding to the roots obtained for the matrix

$$[P] = [W]^{-1}[A] \tag{7.16}$$

where

$[W]^{-1}$ is the inverse of the pooled within groups deviation score cross-products and [A] is the among groups cross-products of deviations of group means from the grand mean (weighted by the group size).

Once the discriminant functions are obtained, they may be used to classify subjects on the basis of their continuous variables. The number of functions to be applied to each individual's set of X scores will be one less than the number of groups or the number of X variables (whichever is less). Subjects are then classified into the group for which their discriminant score has the highest probability of belonging.

Discriminant function analysis and Multivariate Analysis of Variance results are essentially identical. The Wilk's Lambda statistic, the Rao F statistic and the Bartlett Chi-Squared statistic will yield the same inference regarding significant differences among the groups. The discriminant functions may be used to obtain a plot of the subjects in the discriminant space, that is, the Cartesian (orthogonal) space of the discriminant functions. By examining these plots and the standardized coefficients which contribute the most to each discriminant function, you can determine those variables which appear to best differentiate among the groups.

Cluster Analyses

Theory

Objects or people may form groups on the basis of similarity of scores on one or more variables. For example, students in a school may form groups relatively homogeneous with regard to interests in music, athletics, science, languages, etc. An investigator may not have "a priori" groups but rather, be interested in identifying "natural" groupings based on similar score profiles. The Cluster programs of this chapter provide the capability of combining subjects which have the most similar profile of scores.

Hierarchical Cluster Analysis

This procedure was adapted from the Fortran program provided by Donald J. Veldman in his 1967 book. To begin, the sum of squared differences for each pair of subjects on K variables is calculated. If there are n subjects, there are n * (n-1) / 2 pairings. That pair of subjects yielding the smallest sum of squared differences is then combined using the average of the pair on each variable, forming a new "subject" or group. The process is repeated with a new combination formed each time. Eventually, of course, all subjects are combined into a single group. The decision as to when to stop further clustering is typically based on an "error" estimate which reflects the variability of scores for subjects in groups. As in analysis of variance, the between group variability should be significantly greater than the within group variability, if there are to be significant differences among the groups formed.

When you begin execution of the program, you are asked to identify the variables in your data file that are to be used in the grouping. You are also asked to enter the number of groups at which to begin printing the members within each cluster. This may be any value from the total number of subjects down to 2. In practice, you normally select the value of the "ideal" number of groups you expect or some slightly larger value so you can see the increase in error which occurs as more and more of the groups and subjects are combined into new groups. You may also specify the significance level necessary to end the grouping, for example, the value .05 is frequently used in one-way ANOVA analyses when testing for significance. The value used is in fact referred to the F distribution for an F approximation to a multivariate Wilk's Lambda statistic.

Path Analysis

Theory

Path analysis is a procedure for examining the inter-correlations among a set of variables to see if they are consistent with a model of causation. A causal model is one in which the observed scores (events) of an object are assumed to be directly or indirectly caused by one or more preceding events. For example, entrance to college may be hypothesized to be a result of high achievement in high school. High achievement in high school may be the result of parent expectations and the student's intelligence. Intelligence may be a result of parent intelligence, early nutrition, and early environmental stimulation, etc., etc. . Causing and resultant variables may be described in a set of equations. Using standardized z scores, the above example might be described by the following equations:

(1)	$z_1 = e_1$	Parent intelligence
(2)	$z_2 = P_{21}z_1 + e_2$	Child's nutrition
(3)	$z_3 = P_{31}z_1 + P_{32}z_2 + e_3$	Child's intelligence
(4)	$z_4 = P_{41}z_1 + e_4$	Parent expectations
(5)	$z_5 = P_{53}z_3 + p_{54}z_4 + e_5$	School achievement
(6)	$z_6 = P_{63}z_3 + P_{64}z_4 + P_{65}z_5 + e_6$	College GPA

In the above equations, the P's represent path coefficients measuring the strength of causal effect on the resultant due to the causing variable z. In the above example, z_1 has no causing variable and path coefficient. It is called an exogenous variable and is assumed to have only external causes unknown in this model. The "e" values represent contributions that are external and unknown for each variable. These external causes are assumed to be uncorrelated and dropped from further computations in this model. By substituting the definitions of each z score in a model like the above, the correlation between the observed variables can be expressed as in the following examples:

$$r_{12} = \sum z_1 z_2 / n = P_{21}\sum z_1 z_1 / n = P_{21} \quad (7.17)$$

$$r_{23} = \sum z_2 z_3 / n = P_{31}P_{21} + P_{32} \quad (7.18)$$

etc.

In other words, the correlations are estimated to be the sum of direct path coefficients and products of indirect path coefficients. The path coefficients are estimated by the standardized partial regression coefficients (betas) of each resultant variable on its causing variables. For example, coefficients P_{31} and P_{32} above would be estimated by $ß_{31.2}$ and $ß_{32.1}$ in the multiple regression equation

$$z_3 = ß_{31.2}z_1 + ß_{32.1}z_2 + e_3 \quad (7.19)$$

If the hypothesized causal flow model sufficiently describes the interrelationships among the observed variables, the reproduced correlation matrix using the path coefficients should deviate only by sampling error from the original correlations among the variables.

When you execute the Path Analysis procedure in LazStats, you will be asked to specify the exogenous and endogenous variables in your analysis. The program then asks you to specify, for each resultant (endogenous) variable, the causing variables. In this manner you specify your total path model. The program then completes the number of multiple regression analyses required to estimate the path coefficients, estimate the correlations that would be obtained using the model path coefficients and compare the reproduced correlation matrix with the actual correlations among the variables.

You may discover in your reading that this is but one causal model method. More complex methods include models involving latent variables (such as those identified through factor analysis), correlated errors, adjustments for reliability of the variables, etc. Structural model equations of these types are often analyzed using the LISREL™ package found in commercial packages such as SPSS™ or SAS™.

Factor Analysis

The Linear Model

Factor analysis is based on the procedure for obtaining a new set of uncorrelated (orthogonal) variables, usually fewer in number than the original set, that reproduces the co-variability observed among a set or original variables. Two models are commonly utilized:

1. The principal components model wherein the observed score of an individual i on the j<u>th</u> variable $X_{i,j}$ is given as:

$$X_{i,j} = A_{j,1}S_{i,1} + A_{j,2}S_{i,2} + ...+ A_{j,k}S_{i,k} + C \qquad (7.20)$$

where $A_{j,k}$ is a loading of the k<u>th</u> factor on variable j,
$S_{i,k}$ is the factor score of the i<u>th</u> individual on the k<u>th</u> factor and
C is a constant.

The $A_{j,k}$ loadings are typically least-squares regression coefficients.

2. The common factor model assumes each variable X may contain some unique component of variability among subjects as well as components in common with other variables. The model is:

$$X_{i,j} = A_{j,1}S_{i,1} + ... + A_{j,k}S_{i,k} + A_{j,u}S_{i,u} \qquad (7.21)$$

The above equation may also be expressed in terms of standard z scores as:

$$z_{i,j} = a_{j,1}S_{i,1} + ... + a_{j,k}S_{i,k} + a_{j,u}S_{i,u} \qquad (7.22)$$

Since the average of standard score products for the n cases is the product-moment correlation coefficient, the correlation matrix among the j observed variables may be expressed in matrix form as:

$$\underset{jxj}{[R]} = \underset{jxk}{[F]}\ \underset{kxj}{[F]'} - \underset{jxj}{[U]^2} \qquad \text{(array sizes } k <= j) \qquad (7.23)$$

The matrix [F] is the matrix of factor loadings or correlations of the k theoretical orthogonal variables with the j observed variables. The [U] matrix is a diagonal matrix with unique loadings on the diagonal.

The factor loadings above are the result of calculating the eigenvalues and associated vectors of the characteristic equation:

$$| [R] - [U]^2 - [I] | \qquad (7.24)$$

where the lambda values are eigenvalues (roots) of the equation.

When you execute the Factor Analysis Program in LazStats, you are asked to provide information necessary to complete an analysis of your data file. You enter the name of your file and identify the variables to analyze. If you elect to send output to the printer, be sure the printer is on when you start. You will also be asked to specify the type of analysis to perform. The principle components method, a partial image analysis, a Guttman Image Analysis, a Harris Scaled Image Analysis, a Canonical Factor Analysis or an Alpha Factor Analysis may be elected. Selection of the method depends on the assumptions you make concerning sampling of variables and sampling of subjects as well as the theory on which you view your variables. You may request a rotation of the resulting factors which follows completion of the analysis of the data,. The most common rotation performed is the Varimax rotation. This method rotates the orthogonal factor loadings so that the loadings within each factor are most variable on the average. This tends to produce "simple structure", that is, factors which have very high or very low loadings for the original variables and thus simplifies the interpretation of the resulting factors. One may also elect to perform a Procrustean rotation whereby the obtained factor loadings are rotated to be maximally congruent with another factor loading matrix. This second set of loadings which is entered by the user is typically a set which represents some theoretical structure of the original variables. One might, however, obtain factor loadings for males and females separately and then rotate one solution against the other to see if the structures are highly similar for both sexes.

An Example: We will use the cansas.laz file and analyze the relationship between the three predictor variables weight, waist and pulse with the dependent variables chins, situps and jumps.

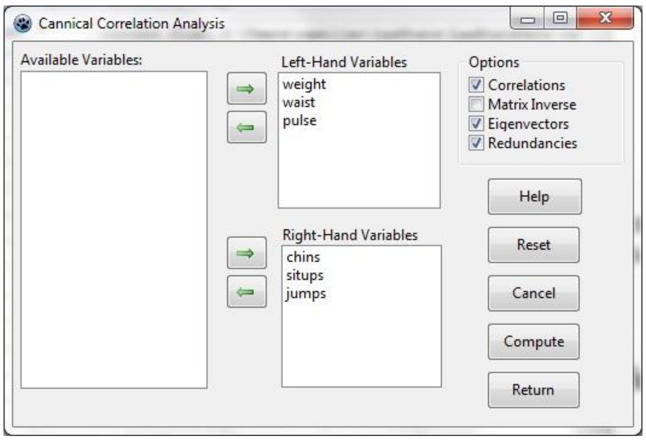


Fig. 7.1 The Cannonical Correlation Dialog

```
The results obtained are:

CANONICAL CORRELATION ANALYSIS

Left Correlation Matrix with   20 cases.

Variables
                 weight        waist        pulse
    weight       1.000        0.870       -0.366
     waist       0.870        1.000       -0.353
     pulse      -0.366       -0.353        1.000

Right Correlation Matrix with   20 cases.

Variables
                  chins       situps        jumps
     chins       1.000        0.696        0.496
    situps       0.696        1.000        0.669
     jumps       0.496        0.669        1.000

Left-Right Correlation Matrix with   20 cases.

Variables
                  chins       situps        jumps
    weight      -0.390       -0.493       -0.226
```

```
     waist      -0.552       -0.646       -0.191
     pulse       0.151        0.225        0.035

Right Inverse x Right-Left Matrix with   20 cases.

Variables
                 weight        waist        pulse
     chins      -0.102       -0.226        0.001
    situps      -0.552       -0.788        0.365
     jumps       0.193        0.448       -0.210

Left Inverse x Left-Right Matrix with   20 cases.

Variables
                  chins       situps        jumps
    weight       0.368        0.287       -0.259
     waist      -0.882       -0.890        0.015
     pulse      -0.026        0.016       -0.055

Canonical Function with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
    Var. 1       0.162        0.172        0.023
    Var. 2       0.482        0.549        0.111
    Var. 3      -0.318       -0.346       -0.032

Trace of the matrix:=    0.6785
Percent of trace extracted:   100.0000

   Canonical R   Root  % Trace   Chi-Sqr    D.F.    Prob.
 1   0.795608    0.633  93.295   16.255       9    0.062
 2   0.200556    0.040   5.928    0.718       4    0.949
 3   0.072570    0.005   0.776    0.082       1    0.775

Overall Tests of Significance:
         Statistic      Approx. Stat.   Value   D.F.  Prob.>Value
Wilk's Lambda          Chi-Squared    17.3037    9   0.0442
Hotelling-Lawley Trace  F-Test          2.4938  9 38  0.0238
Pillai Trace           F-Test          1.5587  9 48  0.1551
Roys Largest Root      F-Test         10.9233  3 19  0.0002

Eigenvectors with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
    Var. 1       0.210       -0.066        0.051
    Var. 2       0.635        0.022       -0.049
    Var. 3      -0.431        0.188        0.017

Standardized Right Side Weights with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
    weight       0.775       -1.884        0.191
     waist      -1.579        1.181       -0.506
     pulse       0.059       -0.231       -1.051

Standardized Left Side Weights with   20 cases.
```

```
Variables
                 Var. 1       Var. 2       Var. 3
     chins        0.349       -0.376        1.297
    situps        1.054        0.123       -1.237
     jumps       -0.716        1.062        0.419

Raw Right Side Weights with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
    weight        0.031       -0.076        0.008
     waist       -0.493        0.369       -0.158
     pulse        0.008       -0.032       -0.146

Raw Left Side Weights with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
     chins        0.066       -0.071        0.245
    situps        0.017        0.002       -0.020
     jumps       -0.014        0.021        0.008

Right Side Correlations with Function with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
    weight       -0.621       -0.772        0.135
     waist       -0.925       -0.378        0.031
     pulse        0.333        0.041       -0.942

Left Side Correlations with Function with   20 cases.

Variables
                 Var. 1       Var. 2       Var. 3
     chins        0.728        0.237        0.644
    situps        0.818        0.573       -0.054
     jumps        0.162        0.959        0.234

Redundancy Analysis for Right Side Variables

            Variance Prop.    Redundancy
        1     0.45080        0.28535
        2     0.24698        0.00993
        3     0.30222        0.00159

Redundancy Analysis for Left Side Variables
            Variance Prop.    Redundancy
        1     0.40814        0.25835
        2     0.43449        0.01748
        3     0.15737        0.00083
```

Interpreting The Standardized Canonical Coefficients.

The elements of [V] represent the standardized weights obtained from the characteristic equation. These elements are the coefficients with which to weight each of the standard (z) scores in our equation (1) above.

Typically, these weights are presented in two parts:

a. The coefficients corresponding to each root are presented as column vectors for the left-hand weights and
b. the coefficients corresponding to each root are presented as column vectors for the right-hand weights.

Structure Coefficents.

In addition to the standardized canonical coefficients, it is useful to obtain what are called structure coefficients. Structure coefficients are the correlations of the left-hand variables with the left-hand composite score (function) and the correlations of the right-hand variables with the right-hand function.

We may also be interested in obtaining and interpreting the correlations of the left-hand function with individual variables of the right-hand and also the correlation of the individual left-hand variables with the right-hand function (for each linear equation).

Discriminant Function / MANOVA

An Example

We will use the file labeled ManoDiscrim.LAZ for our example. A file of the same name (or a .tab file) should be in your directory. Load the file and then click on the Statistics / Multivariate / Discriminant Function option. You should see the form below completed for a discriminant function analysis:

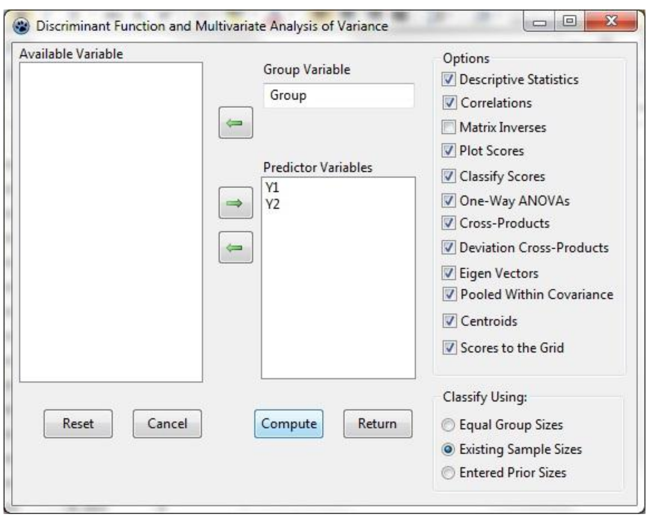


Fig. 7.2 Discriminant Function Analysis Form

You will notice we have asked for all options and have specified that classification use the a priori (sample) sizes for classification. When you click the Compute button, the following results are obtained:

```
MULTIVARIATE ANOVA / DISCRIMINANT FUNCTION
Reference: Multiple Regression in Behavioral Research
Elazar J. Pedhazur, 1997, Chapters 20-21
Harcourt Brace College Publishers

Total Cases := 15, Number of Groups := 3


SUM OF CROSS-PRODUCTS for Group 1, N = 5 with    5 cases.


Variables
                    Y1           Y2
```

```
       Y1      111.000      194.000
       Y2      194.000      343.000



WITHIN GROUP SUM OF DEVIATION CROSS-PRODUCTS Group 1, N := 5 with    5 cases.


Variables
                    Y1           Y2
       Y1        5.200        5.400
       Y2        5.400        6.800




MEANS FOR GROUP 1, N := 5 with    5 valid cases.

Variables          Y1           Y2
                 4.600        8.200


VARIANCES FOR GROUP 1 with    5 valid cases.

Variables          Y1           Y2
                 1.300        1.700


STANDARD DEVIATIONS FOR GROUP 1 with    5 valid cases.

Variables          Y1           Y2
                 1.140        1.304
SUM OF CROSS-PRODUCTS for Group 2, N = 5 with    5 cases.


Variables
                    Y1           Y2
       Y1      129.000      169.000
       Y2      169.000      223.000



WITHIN GROUP SUM OF DEVIATION CROSS-PRODUCTS Group 2, N := 5 with    5 cases.


Variables
                    Y1           Y2
       Y1        4.000        4.000
       Y2        4.000        5.200




MEANS FOR GROUP 2, N := 5 with    5 valid cases.

Variables          Y1           Y2
                 5.000        6.600


VARIANCES FOR GROUP 2 with    5 valid cases.

Variables          Y1           Y2
                 1.000        1.300


STANDARD DEVIATIONS FOR GROUP 2 with    5 valid cases.

Variables          Y1           Y2
1.0      1.140

SUM OF CROSS-PRODUCTS for Group 3, N = 5 with    5 cases.
```

```
Variables
                    Y1           Y2
       Y1      195.000      196.000
       Y2      196.000      199.000



WITHIN GROUP SUM OF DEVIATION CROSS-PRODUCTS Group 3, N := 5 with    5 cases.


Variables
                    Y1           Y2
       Y1        2.800        3.800
       Y2        3.800        6.800



MEANS FOR GROUP 3, N := 5 with    5 valid cases.

Variables          Y1           Y2
                6.200        6.200


VARIANCES FOR GROUP 3 with    5 valid cases.

Variables          Y1           Y2
                0.700        1.700


STANDARD DEVIATIONS FOR GROUP 3 with    5 valid cases.

Variables          Y1           Y2
                0.837        1.304

TOTAL SUM OF CROSS-PRODUCTS with   15 cases.


Variables
                    Y1           Y2
       Y1      435.000      559.000
       Y2      559.000      765.000
TOTAL SUM OF DEVIATION CROSS-PRODUCTS with   15 cases.


Variables
                    Y1           Y2
       Y1       18.933        6.000
       Y2        6.000       30.000

MEANS with   15 valid cases.

Variables          Y1           Y2
                5.267        7.000


VARIANCES with   15 valid cases.

Variables          Y1           Y2
                1.352        2.143


STANDARD DEVIATIONS with   15 valid cases.

Variables          Y1           Y2
                1.163        1.464

BETWEEN GROUPS SUM OF DEV. CPs with   15 cases.


Variables
                    Y1           Y2
```

```
       Y1        6.933       -7.200
       Y2       -7.200       11.200
UNIVARIATE ANOVA FOR VARIABLE Y1
SOURCE    DF       SS        MS        F         PROB > F
BETWEEN     2     6.933     3.467     3.467     0.065
ERROR      12    12.000     1.000
TOTAL      14    18.933

UNIVARIATE ANOVA FOR VARIABLE Y2
SOURCE    DF       SS        MS        F         PROB > F
BETWEEN     2    11.200     5.600     3.574     0.061
ERROR      12    18.800     1.567
TOTAL      14    30.000
Inv. of Pooled Within Dev. CPs Matrix with   15 cases.


Variables
                    Y1           Y2
       Y1        0.366       -0.257
       Y2       -0.257        0.234
Number of roots extracted := 2
Percent of trace extracted :=   100.0000
Roots of the W inverse time B Matrix

No.       Root   Proportion   Canonical R    Chi-Squared  D.F.    Prob.
 1      8.7985   0.9935        0.9476          25.7156     4     0.000
 2      0.0571   0.0065        0.2325           0.6111     1     0.434

Eigenvectors of the W inverse x B Matrix with   15 cases.


Variables
                      1            2
       Y1       -2.316        0.188
       Y2        1.853        0.148

Pooled Within-Groups Covariance Matrix with   15 cases.


Variables
                    Y1           Y2
       Y1        1.000        1.100
       Y2        1.100        1.567


Total Covariance Matrix with   15 cases.


Variables
                    Y1           Y2
       Y1        1.352        0.429
       Y2        0.429        2.143
```

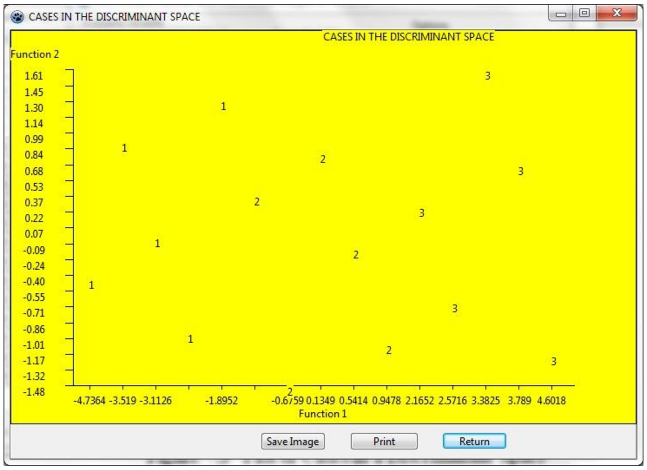


Fig. 7.3   Plot of Cases in a Discriminant Space

```
Raw Function Coeff.s from Pooled Cov. with   15 cases.


Variables
                      1             2
       Y1       -2.030         0.520
       Y2        1.624         0.409




Raw Discriminant Function Constants with   15 valid cases.

Variables            1             2
                -0.674        -5.601
Fisher Discriminant Functions
Group   1 Constant := -24.402
Variable  Coefficient
     1     -5.084
     2      8.804

Group   2 Constant := -14.196
Variable  Coefficient
     1      1.607
     2      3.084

Group   3 Constant := -19.759
Variable  Coefficient
     1      8.112
     2     -1.738
CLASSIFICATION OF CASES
SUBJECT ACTUAL HIGH  PROBABILITY   SEC.D HIGH   DISCRIM
ID NO.  GROUP  IN    GROUP P(G/D)  GROUP P(G/D) SCORE
  1       1      1      0.9999       2    0.0001  4.6019
                                                 -1.1792
  2       1      1      0.9554       2    0.0446  2.5716
                                                 -0.6590
  3       1      1      0.8903       2    0.1097  2.1652
                                                  0.2699
  4       1      1      0.9996       2    0.0004  3.7890
                                                  0.6786
  5       1      1      0.9989       2    0.0011  3.3826
                                                  1.6075
  6       2      2      0.9746       3    0.0252 -0.6760
                                                 -1.4763
  7       2      2      0.9341       1    0.0657  0.9478
                                                 -1.0676
  8       2      2      0.9730       1    0.0259  0.5414
                                                 -0.1387
  9       2      2      0.5724       3    0.4276 -1.4888
                                                  0.3815
 10       2      2      0.9842       1    0.0099  0.1350
```

```
                                              0.7902
 11      3      3      0.9452        2   0.0548 -2.7062
                                             -0.9560
 12      3      3      0.9999        2   0.0001 -4.7365
                                             -0.4358
 13      3      3      0.9893        2   0.0107 -3.1126
                                             -0.0271
 14      3      3      0.9980        2   0.0020 -3.5191
                                              0.9018
 15      3      3      0.8007        2   0.1993 -1.8953
                                              1.3104

CLASSIFICATION TABLE


                        PREDICTED GROUP
Variables
                     1            2            3        TOTAL
         1           5            0            0            5
         2           0            5            0            5
         3           0            0            5            5
     TOTAL           5            5            5           15
Standardized Coeff. from Pooled Cov. with   15 cases.


Variables
                      1            2
       Y1       -2.030        0.520
       Y2        2.032        0.511

Centroids with   15 cases.


Variables
                      1            2
         1       3.302        0.144
         2      -0.108       -0.302
         3      -3.194        0.159
Raw Coefficients from Total Cov. with   15 cases.


Variables
                      1            2
       Y1       -0.701        0.547
       Y2        0.560        0.429




Raw Discriminant Function Constants with   15 valid cases.

Variables            1            2
                -0.674       -5.601


Standardized Coeff.s from Total Cov. with   15 cases.


Variables
                      1            2
       Y1       -0.815        0.636
       Y2        0.820        0.628
Total Correlation Matrix with   15 cases.


Variables
                    Y1           Y2
       Y1        1.000        0.252
       Y2        0.252        1.000



Corr.s Between Variables and Functions with   15 cases.
```

```
Variables
                      1             2
       Y1       -0.608        0.794
       Y2        0.615        0.788



Wilk's Lambda =     0.0965.
F =    12.2013 with D.F.     4 and    22 .  Prob > F = 0.0000
Bartlett Chi-Squared =    26.8845 with 4 D.F. and prob. = 0.0000
Pillai Trace =     0.9520
```

You will notice that we have obtained cross-products and deviation cross-products for each group as well as the combined between and within groups as well as descriptive statistics (means, variances, standard deviations.) Two roots were obtained, the first significant at the 0.05 level using a chi-square test. The one-way analyses of variances completed for each continuous variable were not significant at the 0.05 level which demonstrates that a multivariate analysis may identify group differences not caught by individual variable analysis. The discriminant functions can be used to plot the group subjects in the (orthogonal) space of the functions. If you examine the plot you can see that the individuals in the three groups analyzed are easily separated using just the first discriminant function (the horizontal axis.) Raw and standardized coefficients for the discriminant functions are presented as well as Fisher's discriminant functions for each group. The latter are used to classify the subjects and the classifications are shown along with a table which summarizes the classifications. Note that in this example, all cases are correctly classified. Certainly, a cross-validation of the functions for classification would likely encounter some errors of classification. Since we asked that the discriminant scores be placed in the data grid, the last Fig. shows the data grid with the Fisher discriminant scores saved as two new variables.

Cluster Analyses

To demonstrate the Hierarchical Clustering program, the data to be analyzed is the one labeled cansas.LAZ. You will see the form below with specifications for the grouping:

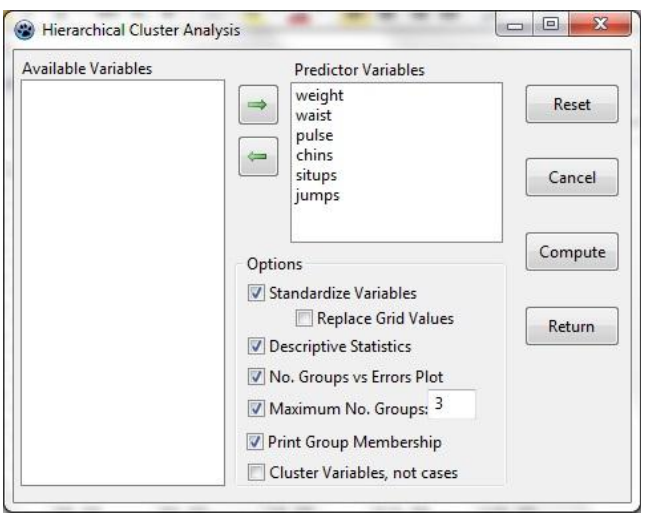


Fig. 7.4 Specifications fo the Hierarchical Cluster Analysis

Results for the hierarchical analysis that you would obtain after clicking the Compute button are presented below:

```
Hierarchical Cluster Analysis

Number of objects to cluster := 20 on 6 variables.
```

```
Variable Means with   20 valid cases.

Variables       weight       waist       pulse       chins      situps
               178.600      35.400      56.100       9.450     145.550

Variables        jumps
                70.300


Variable Variances with   20 valid cases.

Variables       weight       waist       pulse       chins      situps
               609.621      10.253      51.989      27.945    3914.576

Variables        jumps
              2629.379


Variable Standard Deviations with   20 valid cases.

Variables       weight       waist       pulse       chins      situps
                24.691       3.202       7.210       5.286      62.567

Variables        jumps
                51.277

19 groups after combining group 2 (n := 1 ) and group 6 (n := 1) error :=   0.386

18 groups after combining group 18 (n := 1 ) and group 19 (n := 1) error :=   0.387

17 groups after combining group 12 (n := 1 ) and group 18 (n := 2) error :=   0.556

16 groups after combining group 2 (n := 2 ) and group 17 (n := 1) error :=   0.663

15 groups after combining group 4 (n := 1 ) and group 8 (n := 1) error :=   0.805

14 groups after combining group 5 (n := 1 ) and group 11 (n := 1) error :=   1.050

13 groups after combining group 3 (n := 1 ) and group 7 (n := 1) error :=   1.345

12 groups after combining group 2 (n := 3 ) and group 15 (n := 1) error :=   1.402

11 groups after combining group 1 (n := 1 ) and group 2 (n := 4) error :=   1.489

10 groups after combining group 12 (n := 3 ) and group 13 (n := 1) error :=   2.128
Group 1 (n := 5)
  Object := 0
  Object := 1
  Object := 5
  Object := 14
  Object := 16
Group 3 (n := 2)
  Object := 2
  Object := 6
Group 4 (n := 2)
  Object := 3
  Object := 7
Group 5 (n := 2)
  Object := 4
  Object := 10
Group 9 (n := 1)
  Object := 8
Group 10 (n := 1)
  Object := 9
Group 12 (n := 4)
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 14 (n := 1)
  Object := 13
Group 16 (n := 1)
  Object := 15
Group 20 (n := 1)
```

```
  Object := 19

9 groups after combining group 4 (n := 2 ) and group 20 (n := 1) error :=   2.721
Group 1 (n := 5)
  Object := 0
  Object := 1
  Object := 5
  Object := 14
  Object := 16
Group 3 (n := 2)
  Object := 2
  Object := 6
Group 4 (n := 3)
  Object := 3
  Object := 7
  Object := 19
Group 5 (n := 2)
  Object := 4
  Object := 10
Group 9 (n := 1)
  Object := 8
Group 10 (n := 1)
  Object := 9
Group 12 (n := 4)
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 14 (n := 1)
  Object := 13
Group 16 (n := 1)
  Object := 15

8 groups after combining group 3 (n := 2 ) and group 16 (n := 1) error :=   3.151
Group 1 (n := 5)
  Object := 0
  Object := 1
  Object := 5
  Object := 14
  Object := 16
Group 3 (n := 3)
  Object := 2
  Object := 6
  Object := 15
Group 4 (n := 3)
  Object := 3
  Object := 7
  Object := 19
Group 5 (n := 2)
  Object := 4
  Object := 10
Group 9 (n := 1)
  Object := 8
Group 10 (n := 1)
  Object := 9
Group 12 (n := 4)
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 14 (n := 1)
  Object := 13

7 groups after combining group 4 (n := 3 ) and group 9 (n := 1) error :=   6.111
Group 1 (n := 5)
  Object := 0
  Object := 1
  Object := 5
  Object := 14
  Object := 16
Group 3 (n := 3)
  Object := 2
  Object := 6
  Object := 15
```

```
Group 4 (n := 4)
  Object := 3
  Object := 7
  Object := 8
  Object := 19
Group 5 (n := 2)
  Object := 4
  Object := 10
Group 10 (n := 1)
  Object := 9
Group 12 (n := 4)
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 14 (n := 1)
  Object := 13

6 groups after combining group 5 (n := 2 ) and group 12 (n := 4) error :=   6.180
Group 1 (n := 5)
  Object := 0
  Object := 1
  Object := 5
  Object := 14
  Object := 16
Group 3 (n := 3)
  Object := 2
  Object := 6
  Object := 15
Group 4 (n := 4)
  Object := 3
  Object := 7
  Object := 8
  Object := 19
Group 5 (n := 6)
  Object := 4
  Object := 10
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 10 (n := 1)
  Object := 9
Group 14 (n := 1)
  Object := 13

5 groups after combining group 1 (n := 5 ) and group 3 (n := 3) error :=   7.617
Group 1 (n := 8)
  Object := 0
  Object := 1
  Object := 2
  Object := 5
  Object := 6
  Object := 14
  Object := 15
  Object := 16
Group 4 (n := 4)
  Object := 3
  Object := 7
  Object := 8
  Object := 19
Group 5 (n := 6)
  Object := 4
  Object := 10
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 10 (n := 1)
  Object := 9
Group 14 (n := 1)
  Object := 13

4 groups after combining group 5 (n := 6 ) and group 10 (n := 1) error :=  11.027
```

```
Group 1 (n := 8)
  Object := 0
  Object := 1
  Object := 2
  Object := 5
  Object := 6
  Object := 14
  Object := 15
  Object := 16
Group 4 (n := 4)
  Object := 3
  Object := 7
  Object := 8
  Object := 19
Group 5 (n := 7)
  Object := 4
  Object := 9
  Object := 10
  Object := 11
  Object := 12
  Object := 17
  Object := 18
Group 14 (n := 1)
  Object := 13

3 groups after combining group 1 (n := 8 ) and group 14 (n := 1) error :=  13.897
Group 1 (n := 9)
  Object := 0
  Object := 1
  Object := 2
  Object := 5
  Object := 6
  Object := 13
  Object := 14
  Object := 15
  Object := 16
Group 4 (n := 4)
  Object := 3
  Object := 7
  Object := 8
  Object := 19
Group 5 (n := 7)
  Object := 4
  Object := 9
  Object := 10
  Object := 11
  Object := 12
  Object := 17
  Object := 18

2 groups after combining group 4 (n := 4 ) and group 5 (n := 7) error :=  17.198
Group 1 (n := 9)
  Object := 0
  Object := 1
  Object := 2
  Object := 5
  Object := 6
  Object := 13
  Object := 14
  Object := 15
  Object := 16
Group 4 (n := 11)
  Object := 3
  Object := 4
  Object := 7
  Object := 8
  Object := 9
  Object := 10
  Object := 11
  Object := 12
  Object := 17
  Object := 18
  Object := 19
```

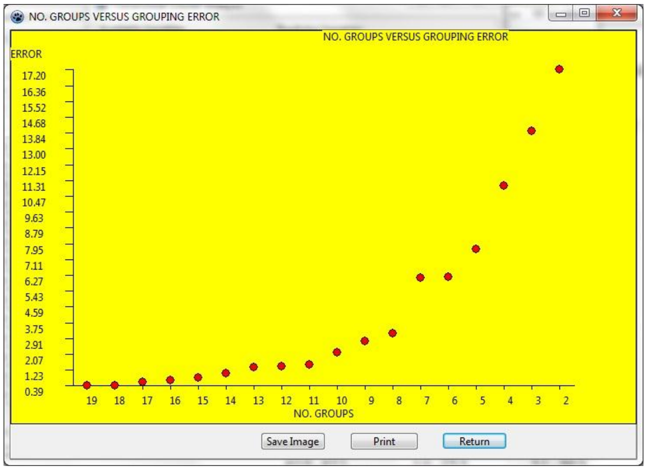


Fig. 7.5 Grouping Errors in Hierarchical Clustering

If you compare the results above with a discriminant analysis on the same data, you will see that the clustering procedure does not necessarily replicate the original groups. Clearly, “nearest neighbor” grouping in Euclidean space does not necessarily result in the same a priori groups from the discriminant analysis.

By examining the increase in error (variance of subjects within the groups) as a function of the number of groups, one can often make some decision about the number of groups they wish to interpret. There is a large increase in error when going from 8 groups down to 7 in this analysis which suggests there are possibly 7 or 8 groups which might be examined. If we had more information on the objects of those groups, we might see a pattern or commonality shared by objects of those groups.

K-Means Clustering Analysis

With this procedure, one first specifies the number of groups to be formed among the objects. The procedure uses a procedure to load each of the k groups with one object in a somewhat random manner. The procedure then iteratively adds or subtracts objects from each group based on an error measure of the distance between the objects in the group. The procedure ends when subsequent iterations do not produce a lower value or the number of iterations has been exceeded.

In this example, we loaded the cansas.LAZ file to group the 20 subjects into four groups. The results may be compared with the other cluster methods of this chapter.

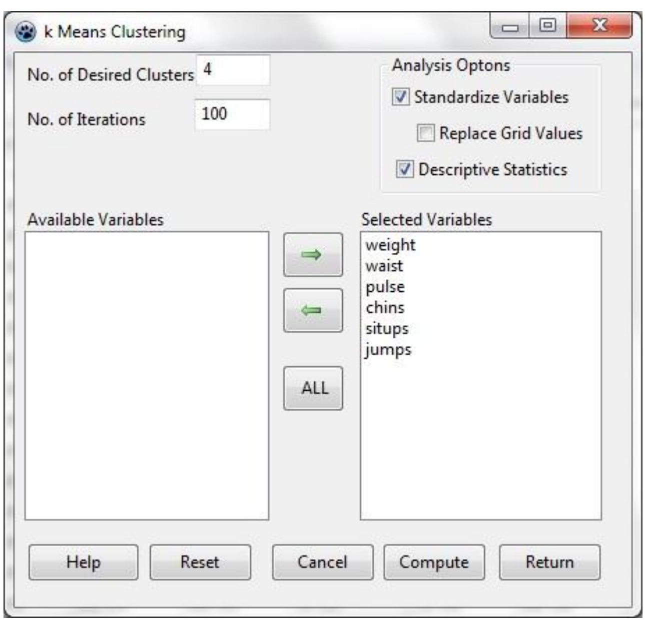


Fig. 7.6 The K-Means Clustering Form

Results are:

```
K-Means Clustering.  Adapted from AS 136  APPL. STATIST. (1979) VOL.28, NO.1

File := C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ
No. Cases := 20, No. Variables := 6, No. Clusters := 4

Mean :=  178.600, Std.Dev. :=  609.621 for weight
Mean :=   35.400, Std.Dev. :=   10.253 for waist
Mean :=   56.100, Std.Dev. :=   51.989 for pulse
Mean :=    9.450, Std.Dev. :=   27.945 for chins
Mean :=  145.550, Std.Dev. := 3914.576 for situps
Mean :=   70.300, Std.Dev. := 2629.379 for jumps

NUMBER OF SUBJECTS IN EACH CLUSTER
Cluster := 1 with 1 cases.
Cluster := 2 with 7 cases.
Cluster := 3 with 9 cases.
Cluster := 4 with 3 cases.

PLACEMENT OF SUBJECTS IN CLUSTERS
CLUSTER SUBJECT
     1     14
     2      2
     2      6
     2      8
     2      1
     2     15
     2     17
     2     20
     3     11
     3     12
     3     13
     3      4
     3      5
     3      9
     3     18
     3     19
     3     10
     4      7
     4     16
     4      3

AVERAGE VARIABLE VALUES BY CLUSTER
               VARIABLES
CLUSTER    1     2     3     4     5     6

     1  0.11  1.03 -0.12 -0.30 -0.02 -0.01
     2 -0.00  0.02 -0.02 -0.19 -0.01 -0.01
     3 -0.02 -0.20  0.01  0.17  0.01  0.01
     4  0.04  0.22  0.05  0.04 -0.00  0.01

WITHIN CLUSTER SUMS OF SQUARES
Cluster 1 :=  0.000
Cluster 2 :=  0.274
Cluster 3 :=  0.406
Cluster 4 :=  0.028
```

Average Linkage Hierarchical Cluster Analysis

This cluster procedure clusters objects based on their similarity (or dissimilarity) as recorded in a data matrix. The correlation among objects is often used as a measure of similarity. In this example, we first loaded the file labeled "cansas.laz". We then "rotated" the data using the rotate function in the Edit menu so that columns represent subjects and rows represent variables. We then used the Correlation procedure (with the option to save the correlation matrix) to obtain the correlation among the 20 subjects as a measure of similarity. We then closed the file. Next, we opened the matrix file we had just saved using the File / Open a Matrix File option. We then clicked on the Analyses / Multivariate / Cluster / Average Linkage option. Shown below is the dialogue box for the analysis:

Fig. 7.7   Average Linkage Dialog

Output of the analysis includes a listing of which objects (groups) are combined at each step followed by a dendogram of the combinations. You can compare this method of clustering subjects with that obtained in the previous analysis.

```
Average Linkage Cluster Analysis.  Adopted from ClusBas by John S. Uebersax

Group   1 is joined by group   3. N is   2 ITER :=   1 SIM :=    193.000
Group   1 is joined by group   5. N is   3 ITER :=   2 SIM :=    131.500
Group   1 is joined by group   6. N is   4 ITER :=   3 SIM :=     73.000
Group   1 is joined by group   2. N is   5 ITER :=   4 SIM :=     36.250
Group   1 is joined by group   4. N is   6 ITER :=   5 SIM :=      7.200

No. of objects :=   6
Matrix defined similarities among objects.
Average Linkage Cluster Analysis.  Adopted from ClusBas by John S. Uebersax

Group   1 is joined by group   3. N is   2 ITER :=   1 SIM :=    193.000
Group   1 is joined by group   5. N is   3 ITER :=   2 SIM :=    131.500
Group   1 is joined by group   6. N is   4 ITER :=   3 SIM :=     73.000
Group   1 is joined by group   2. N is   5 ITER :=   4 SIM :=     36.250
Group   1 is joined by group   4. N is   6 ITER :=   5 SIM :=      7.200

No. of objects :=   6
Matrix defined similarities among objects.


UNIT     1    3    5    6    2    4
STEP     *    *    *    *    *    *
    1    ******    *    *    *    *
           *       *    *    *    *
    2      *********    *    *    *
               *        *    *    *
    3          **********    *    *
                   *         *    *
    4              ***********    *
                        *         *
    5                   ***********
                             *
```

Single Link Clustering

This procedure reads a file of subjects or objects measured on one or more variables. One variable is selected to "link" subjects together into groups. Originally, each subject is a group. Subjects closest together on the measure are combined to form a new group with a score that is the average of the two subjects within the group. This process is repeated until only 1 group remains. You can elect to show each grouping step and the errors of grouping as well as a dendogram of the groupings.

Shown below is a single link cluster of subject “jumps” as found in the cansas.laz file:

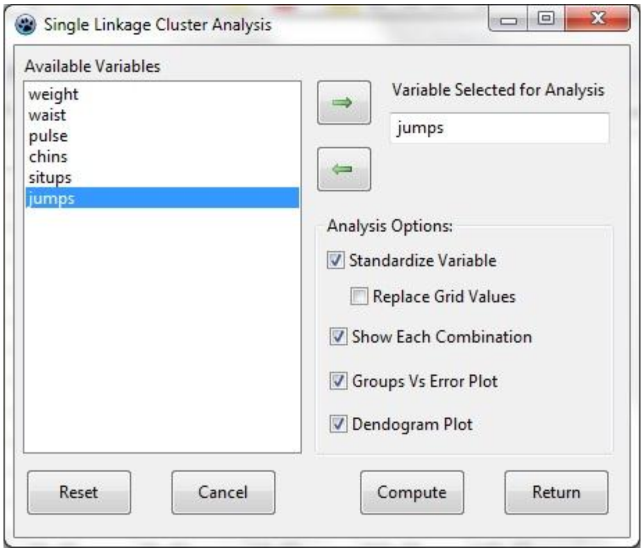


Fig. 7.8 Single Link Clustering Form

```
Single Linkage Clustering by Bill Miller
FILE: C:\Documents and Settings\Owner\My Documents\Projects\CLanguage\data\cansas.LAZ
Variable = jumps
Number of cases = 20
Mean =   70.300, Variance = 2629.379, Std.Dev. =   51.277

GROUP ID  17  15   4   7  11   9   8   6  20  14   5   2   1  19  18   3  13  12  16  10
        (Group 2 is combined with Group 1)

GROUP ID  17  15   4   7  11   9   8   6  20  14   5   2  19  18   3  13  12  16  10
        (Group 9 is combined with Group 8)

GROUP ID  17  15   4   7  11   9   6  20  14   5   2  19  18   3  13  12  16  10
        (Group 7 is combined with Group 11)

GROUP ID  17  15   4   7   9   6  20  14   5   2  19  18   3  13  12  16  10
        (Group 9 is combined with Group 3)

GROUP ID  17  15   4   7   9   6  20  14   5   2  19  18  13  12  16  10
        (Group 2 is combined with Group 18)

GROUP ID  17  15   4   7   9   6  20  14   5   2  19  13  12  16  10
        (Group 6 is combined with Group 20)

GROUP ID  17  15   4   7   9   6  14   5   2  19  13  12  16  10
        (Group 7 is combined with Group 9)

GROUP ID  17  15   4   7   6  14   5   2  19  13  12  16  10
        (Group 7 is combined with Group 13)

GROUP ID  17  15   4   7   6  14   5   2  19  12  16  10
        (Group 12 is combined with Group 16)

GROUP ID  17  15   4   7   6  14   5   2  19  12  10
        (Group 7 is combined with Group 6)

GROUP ID  17  15   4   7  14   5   2  19  12  10
        (Group 15 is combined with Group 4)

GROUP ID  17  15   7  14   5   2  19  12  10
        (Group 15 is combined with Group 7)

GROUP ID  17  15  14   5   2  19  12  10
        (Group 2 is combined with Group 19)

GROUP ID  17  15  14   5   2  12  10
        (Group 14 is combined with Group 5)
```

```
GROUP ID  17  15  14   2  12  10
        (Group 14 is combined with Group 2)

GROUP ID  17  15  14  12  10
        (Group 15 is combined with Group 12)

GROUP ID  17  15  14  10
        (Group 15 is combined with Group 14)

GROUP ID  17  15  10
        (Group 17 is combined with Group 15)

GROUP ID  17  10
        (Group 17 is combined with Group 10)

GROUP ID  17

GROUPING STEP    ERROR
       1              0.000
       2              0.000
       3              0.000
       4              0.008
       5              0.012
       6              0.020
       7              0.035
       8              0.064
       9              0.098
      10              0.102
      11              0.117
      12              0.115
      13              0.142
      14              0.156
      15              0.194
      16              0.270
      17              0.565
      18              1.387
      19              3.314

SCATTERPLOT - Plot of Error vs No. of Groups

                                                         Size of Error
.                             |                             |-  3.48
                              |                             |-  3.31
                              |                             |-  3.15
                              |                             |-  2.98
                              |                             |-  2.82
                              |                             |-  2.65
                              |                             |-  2.49
                              |                             |-  2.32
                              |                             |-  2.15
                              |                             |-  1.99
------------------------------------------------------------|-  1.82
.                             |                             |-  1.66
                              |                             |-  1.49
                              |                             |-  1.33
                              |                             |-  1.16
                              |                             |-  0.99
   .                          |                             |-  0.83
                              |                             |-  0.66
      .   .                   |                             |-  0.50
             .  .   .  .  .   *  .  .   .  .  .             |-  0.33
_______________________________________________________________
   |     |     |     |     |     |     |     |     |     |
No. of Groups
  2.00  4.00  6.00  8.00 10.00 12.00 14.00 16.00 18.00 20.00
```

```
PART 1 OUTPUT
UNIT    17   15    4    7   11    9    8    6   20   14    5    2    1   19   18
STEP     *    *    *    *    *    *    *    *    *    *    *    *    *    *    *
  1      *    *    *    *    *    *    *    *    *    *    *    ******    *    *
         *    *    *    *    *    *    *    *    *    *    *      *       *    *
  2      *    *    *    *    *    ******    *    *    *    *      *       *    *
         *    *    *    *    *      *       *    *    *    *      *       *    *
  3      *    *    *    ******      *       *    *    *    *      *       *    *
         *    *    *      *         *       *    *    *    *      *       *    *
  4      *    *    *      *         *********************************************
         *    *    *      *                 *    *    *    **     *       *    *
  5      *    *    *      *                 *    *    *    **     **************
         *    *    *      *                 *    *    *    **           * *
  6      *    *    *      *                 ******    *    **           * *
         *    *    *      *                   *       *    **           * *
  7      *    *    *      ***********************************           * *
         *    *    *                       *  *       *    *            * *
  8      *    *    *                       **************************************
         *    *    *                          *       *    *      *     * *
  9      *    *    *                          *       *    *      *     * *
         *    *    *                          *       *    *      *     * *
 10      *    *    *                                  * *  *            * *
         *    *    *                                  * *  *            * *
 11      *    ******                                  * *  *            * *
         *      *                                     * *  *            * *
 12      *      *****************************************  *            * *
         *                          *                 *    *            * *
 13      *                          *                 *    *            ***
         *                          *                 *    *             *
 14      *                          *                 ******             *
         *                          *                   *                *
 15      *                          *                   ******************
         *                          *                           *
 16      *                          *********************************************
         *                                                      * *
 17      *                                                       *
         *                                                       *
 18      *********************************************************
                                     *
 19                                  ********************************************
                                                                      *
 20                                                                   *
                                                                      *
```

```
PART 2 OUTPUT
   3   13   12   16   10
   *    *    *    *    *
   *    *    *    *    *
   *    *    *    *    *
   *    *    *    *    *
   *    *    *    *    *
   *    *    *    *    *
   *    *    *    *    *
****    *    *    *    *
        *    *    *    *
        *    *    *    *
        *    *    *    *
        *    *    *    *
        *    *    *    *
        *    *    *    *
        *    *    *    *
*********    *    *    *
             *    *    *
             ******    *
               *       *
               *       *
               *       *
               *       *
               *       *
               *       *
               *       *
               *       *
               *       *
```

```
              *       *
              *       *
              *       *
              *       *
***************       *
                      *
                      *
                      *
                      *
                      *
***********************
```

Path Analysis

Example of a Path Analysis

In this example we will use the file CANSAS.LAZ. The user begins by selecting the Path Analysis option of the Statistics / Multivariate menu. In the Fig. below we have selected all variables to analyze and have entered our first path indicating that waist size is "caused" by weight:

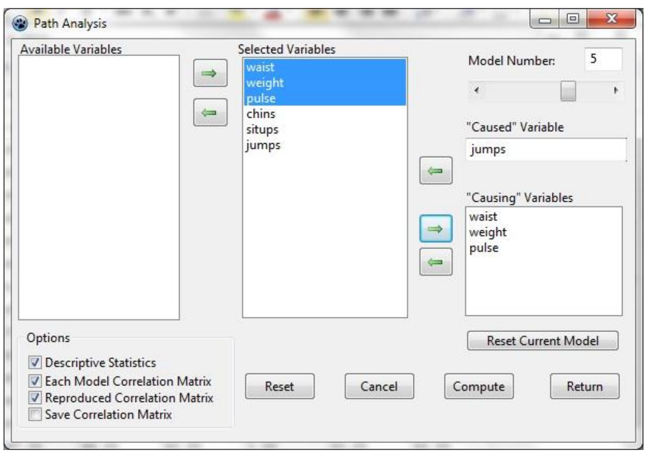


Fig. 7.9 Path Analysis Form

We will also hypothesize that pulse rate is "caused" by weight, chin-ups are "caused" by weight, waist and pulse, that the number of sit-ups is "caused" by weight, waist and pulse and that jumps are "caused" by weight, waist and pulse. Each time we enter a new causal relationship we click the scroll bar to move to a new model number prior to entering the "caused" and "causing" variables. Once we have entered each model, we then click on the Compute button. Note we have elected to print descriptive statistics, each models correlation matrix, and the reproduced correlation matrix which will be our measure of how well the models "fit" the data. The results are shown below:

```
PATH ANALYSIS RESULTS

CAUSED VARIABLE: waist
     Causing Variables:
    weight
CAUSED VARIABLE: pulse
     Causing Variables:
    weight
CAUSED VARIABLE: chins
     Causing Variables:
    weight
    waist
    pulse
CAUSED VARIABLE: situps
     Causing Variables:
```

```
    weight
    waist
    pulse
CAUSED VARIABLE: jumps
     Causing Variables:
    weight
    waist
    pulse


Correlation Matrix with   20 valid cases.

Variables
                weight        waist        pulse        chins       situps
   weight        1.000        0.870       -0.366       -0.390       -0.493
    waist        0.870        1.000       -0.353       -0.552       -0.646
    pulse       -0.366       -0.353        1.000        0.151        0.225
    chins       -0.390       -0.552        0.151        1.000        0.696
   situps       -0.493       -0.646        0.225        0.696        1.000
     jumps      -0.226       -0.191        0.035        0.496        0.669


Variables
                  jumps
   weight       -0.226
    waist       -0.191
    pulse        0.035
    chins        0.496
   situps        0.669
     jumps       1.000


MEANS with   20 valid cases.

Variables      weight        waist        pulse        chins       situps
              178.600       35.400       56.100        9.450      145.550

Variables        jumps
               70.300

VARIANCES with   20 valid cases.

Variables      weight        waist        pulse        chins       situps
              609.621       10.253       51.989       27.945     3914.576

Variables        jumps
             2629.379


STANDARD DEVIATIONS with   20 valid cases.

Variables      weight        waist        pulse        chins       situps
               24.691        3.202        7.210        5.286       62.567

Variables        jumps
               51.277

Dependent Variable = waist


Correlation Matrix with   20 valid cases.


Variables
                weight        waist
   weight        1.000        0.870
    waist        0.870        1.000




MEANS with   20 valid cases.

Variables      weight        waist
```

```
               178.600       35.400


VARIANCES with   20 valid cases.

Variables      weight        waist
               609.621       10.253


STANDARD DEVIATIONS with   20 valid cases.

Variables      weight        waist
                24.691        3.202

Dependent Variable = waist

       R        R2         F     Prob.>F  DF1  DF2
   0.870     0.757    56.173     0.000    1   18
Adjusted R Squared = 0.744

Std. Error of Estimate =      1.621

Variable       Beta      B         Std.Error t         Prob.>t
   weight      0.870     0.113     0.015     7.495     0.000

Constant =     15.244

Dependent Variable = pulse


Correlation Matrix with   20 valid cases.


Variables
                weight        pulse
   weight        1.000       -0.366
    pulse       -0.366        1.000


MEANS with   20 valid cases.

Variables      weight        pulse
               178.600       56.100


VARIANCES with   20 valid cases.

Variables      weight        pulse
               609.621       51.989


STANDARD DEVIATIONS with   20 valid cases.

Variables      weight        pulse
                24.691        7.210

Dependent Variable = pulse

       R        R2         F     Prob.>F  DF1  DF2
   0.366     0.134     2.780     0.113    1   18
Adjusted R Squared = 0.086

Std. Error of Estimate =      6.895

Variable       Beta      B         Std.Error t         Prob.>t
   weight     -0.366    -0.107     0.064    -1.667     0.113

Constant =     75.177

Dependent Variable = chins
```

```
Correlation Matrix with   20 valid cases.


Variables
                weight        waist        pulse        chins
   weight        1.000        0.870       -0.366       -0.390
    waist        0.870        1.000       -0.353       -0.552
    pulse       -0.366       -0.353        1.000        0.151
    chins       -0.390       -0.552        0.151        1.000




MEANS with   20 valid cases.

Variables      weight        waist        pulse        chins
              178.600       35.400       56.100        9.450


VARIANCES with   20 valid cases.

Variables      weight        waist        pulse        chins
              609.621       10.253       51.989       27.945


STANDARD DEVIATIONS with   20 valid cases.

Variables      weight        waist        pulse        chins
               24.691        3.202        7.210        5.286

Dependent Variable = chins

       R        R2         F     Prob.>F  DF1  DF2
   0.583     0.340     2.742     0.077    3   16
Adjusted R Squared = 0.216

Std. Error of Estimate =      4.681

Variable       Beta      B         Std.Error t         Prob.>t
   weight      0.368     0.079     0.089     0.886     0.389
    waist     -0.882    -1.456     0.683    -2.132     0.049
    pulse     -0.026    -0.019     0.160    -0.118     0.907

Constant =     47.968

Dependent Variable = situps


Correlation Matrix with   20 valid cases.


Variables
                weight        waist        pulse       situps
   weight        1.000        0.870       -0.366       -0.493
    waist        0.870        1.000       -0.353       -0.646
    pulse       -0.366       -0.353        1.000        0.225
   situps       -0.493       -0.646        0.225        1.000




MEANS with   20 valid cases.

Variables      weight        waist        pulse       situps
              178.600       35.400       56.100      145.550


VARIANCES with   20 valid cases.

Variables      weight        waist        pulse       situps
              609.621       10.253       51.989     3914.576
STANDARD DEVIATIONS with   20 valid cases.
```

Variables	weight	waist	pulse	situps
	24.691	3.202	7.210	62.567

Dependent Variable = situps

R	R2	F	Prob.>F	DF1	DF2
0.661	0.436	4.131	0.024	3	16

Adjusted R Squared = 0.331

Std. Error of Estimate = 51.181

Variable	Beta	B	Std.Error	t	Prob.>t
weight	0.287	0.728	0.973	0.748	0.466
waist	-0.890	-17.387	7.465	-2.329	0.033
pulse	0.016	0.139	1.755	0.079	0.938

Constant = 623.282

Dependent Variable = jumps

Correlation Matrix with 20 valid cases.

Variables	weight	waist	pulse	jumps
weight	1.000	0.870	-0.366	-0.226
waist	0.870	1.000	-0.353	-0.191
pulse	-0.366	-0.353	1.000	0.035
jumps	-0.226	-0.191	0.035	1.000

MEANS with 20 valid cases.

Variables	weight	waist	pulse	jumps
	178.600	35.400	56.100	70.300

VARIANCES with 20 valid cases.

Variables	weight	waist	pulse	jumps
	609.621	10.253	51.989	2629.379

STANDARD DEVIATIONS with 20 valid cases.

Variables	weight	waist	pulse	jumps
	24.691	3.202	7.210	51.277

Dependent Variable = jumps

R	R2	F	Prob.>F	DF1	DF2
0.232	0.054	0.304	0.822	3	16

Adjusted R Squared = -0.123

Std. Error of Estimate = 54.351

Variable	Beta	B	Std.Error	t	Prob.>t
weight	-0.259	-0.538	1.034	-0.520	0.610
waist	0.015	0.234	7.928	0.029	0.977
pulse	-0.055	-0.389	1.863	-0.209	0.837

Constant = 179.887

Matrix of Path Coefficients with 20 valid cases.

Variables

	weight	waist	pulse	chins	situps
weight	0.000	0.870	-0.366	0.368	0.287
waist	0.870	0.000	0.000	-0.882	-0.890
pulse	-0.366	0.000	0.000	-0.026	0.016
chins	0.368	-0.882	-0.026	0.000	0.000
situps	0.287	-0.890	0.016	0.000	0.000
jumps	-0.259	0.015	-0.055	0.000	0.000

Variables

	jumps
weight	-0.259
waist	0.015
pulse	-0.055
chins	0.000
situps	0.000
jumps	0.000

SUMMARY OF CAUSAL MODELS

Var. Caused	Causing Var.	Path Coefficient
waist	weight	0.870
pulse	weight	-0.366
chins	weight	0.368
chins	waist	-0.882
chins	pulse	-0.026
situps	weight	0.287
situps	waist	-0.890
situps	pulse	0.016
jumps	weight	-0.259
jumps	waist	0.015
jumps	pulse	-0.055

Reproduced Correlation Matrix with 20 valid cases.

Variables

	weight	waist	pulse	chins	situps
weight	1.000	0.870	-0.366	-0.390	-0.493
waist	0.870	1.000	-0.318	-0.553	-0.645
pulse	-0.366	-0.318	1.000	0.120	0.194
chins	-0.390	-0.553	0.120	1.000	0.382
situps	-0.493	-0.645	0.194	0.382	1.000
jumps	-0.226	-0.193	0.035	0.086	0.108

Variables

	jumps
weight	-0.226
waist	-0.193
pulse	0.035
chins	0.086
situps	0.108
jumps	1.000

Average absolute difference between observed and reproduced
coefficients := 0.077
Maximum difference found := 0.562

Data Array of Subject Path z Scores with 20 cases.

Variables

	weight	waist	pulse	chins	situps
Subject 1	0.502	0.437	-0.184	-0.196	0.263
Subject 2	0.421	0.367	-0.154	-0.164	-0.568
Subject 3	0.583	0.508	-0.213	-0.227	-0.712
Subject 4	-0.672	-0.585	0.246	0.262	-0.648
Subject 5	0.421	0.367	-0.154	-0.164	0.151

```
 Subject 6       0.138        0.120       -0.050       -0.054       -0.712
 Subject 7       1.312        1.142       -0.480       -0.511       -0.712
 Subject 8      -0.470       -0.409        0.172        0.183       -0.328
 Subject 9      -0.105       -0.092        0.039        0.041        0.870
Subject 10      -0.996       -0.867        0.364        0.388        1.685
Subject 11      -0.389       -0.338        0.142        0.152       -0.408
Subject 12      -0.510       -0.444        0.187        0.199        1.030
Subject 13      -0.996       -0.867        0.364        0.388        1.110
Subject 14       2.770        2.411       -1.013       -1.080       -1.527
Subject 15       0.583        0.508       -0.213       -0.227       -1.208
Subject 16       0.948        0.825       -0.347       -0.369        1.030
Subject 17      -0.105       -0.092        0.039        0.041       -1.367
Subject 18      -0.875       -0.761        0.320        0.341        1.350
Subject 19      -0.915       -0.797        0.335        0.357        1.270
Subject 20      -1.644       -1.431        0.601        0.641       -0.568

Variables
                  jumps
 Subject 1      -0.114
 Subject 2      -0.095
 Subject 3      -0.132
 Subject 4       0.152
 Subject 5      -0.095
 Subject 6      -0.031
 Subject 7      -0.297
 Subject 8       0.106
 Subject 9       0.024
Subject 10       0.225
Subject 11       0.088
Subject 12       0.115
Subject 13       0.225
Subject 14      -0.627
Subject 15      -0.132
Subject 16      -0.214
Subject 17       0.024
Subject 18       0.198
Subject 19       0.207
Subject 20       0.372
```

We note that pulse is not a particularly important predictor of chin-ups or sit-ups. The largest discrepancy of 0.562 between an original correlation and a correlation reproduced using the path coefficients indicates our model of causation may have been inadequate.

Factor Analysis

The sample factor analysis completed below utilizes a data set labeled cansas.laz as used in the previous path analysis example . The canonical factor analysis method was used andthe varimax rotation method was used.

Shown below is the factor analysis form selected by choosing the factor analysis option under the Statistics / Multivariate menu:

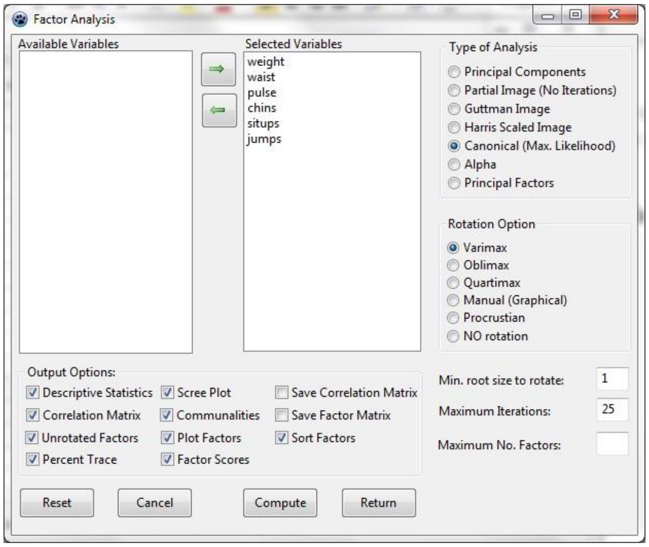


Fig. 7.10   Factor Analysis Dialog

Note the options elected in the above form. The results obtained are shown below:

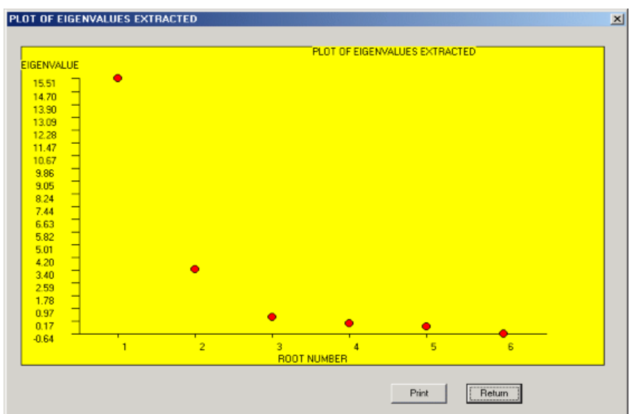


Fig. 7.11   Scree Plot of Eigenvalues

```
Factor Analysis
See Rummel, R.J., Applied Factor Analysis
Northwestern University Press, 1970

Canonical Factor Analysis
Original matrix trace =  18.56
Roots (Eigenvalues) Extracted:
   1 15.512
   2  3.455
   3  0.405
   4  0.010
   5 -0.185
   6 -0.641

Unrotated Factor Loadings

FACTORS with   20 valid cases.


Variables
               Factor 1     Factor 2     Factor 3     Factor 4     Factor 5
   weight        0.858       -0.286        0.157       -0.006        0.000
    waist        0.928       -0.201       -0.066       -0.003        0.000
```

```
    pulse       -0.360        0.149       -0.044       -0.089        0.000
    chins       -0.644       -0.382        0.195        0.009        0.000
   situps       -0.770       -0.472        0.057       -0.009        0.000
     jumps      -0.409       -0.689       -0.222        0.005        0.000


Variables
               Factor 6
   weight        0.000
    waist        0.000
    pulse        0.000
    chins        0.000
   situps        0.000
     jumps       0.000


Percent of Trace In Each Root:
   1 Root := 15.512 Trace := 18.557 Percent :=  83.593
   2 Root :=  3.455 Trace := 18.557 Percent :=  18.621
   3 Root :=  0.405 Trace := 18.557 Percent :=   2.180
   4 Root :=  0.010 Trace := 18.557 Percent :=   0.055
   5 Root := -0.185 Trace := 18.557 Percent :=  -0.995
   6 Root := -0.641 Trace := 18.557 Percent :=  -3.455

COMMUNALITY ESTIMATES
  1 weight      0.844
  2 waist       0.906
  3 pulse       0.162
  4 chins       0.598
  5 situps      0.819
  6 jumps       0.692

Proportion of variance in unrotated factors

  1 48.364
  2 16.475


Communality Estimates as percentages:
  1 81.893
  2 90.153
  3 15.165
  4 56.003
  5 81.607
  6 64.217

Varimax Rotated Loadings with   20 valid cases.

Variables
               Factor 1     Factor 2
   weight       -0.882       -0.201
    waist       -0.898       -0.310
    pulse        0.385        0.059
    chins        0.352        0.660
   situps        0.413        0.803
     jumps      -0.009        0.801

Percent of Variation in Rotated Factors
Factor   1 33.776
Factor   2 31.064

Total Percent of Variance in Factors : 64.840
Communalities as Percentages
  1 for        weight  81.893
  2 for         waist  90.153
  3 for         pulse  15.165
  4 for         chins  56.003
  5 for        situps  81.607
  6 for          jumps 64.217
```

```
             SCATTERPLOT - FACTOR PLOT
Factor 2
 |                             |                              |-  0.95-  1.00
 |                             |                              |-  0.90-  0.95
 |                             |                              |-  0.85-  0.90
 |                             2           1                  |-  0.80-  0.85
 |                             |                              |-  0.75-  0.80
 |                             |                              |-  0.70-  0.75
 |                             |          3                   |-  0.65-  0.70
 |                             |                              |-  0.60-  0.65
 |                             |                              |-  0.55-  0.60
 |                             |                              |-  0.50-  0.55
 |                             |                              |-  0.45-  0.50
 |                             |                              |-  0.40-  0.45
 |                             |                              |-  0.35-  0.40
 |                             |                              |-  0.30-  0.35
 |                             |                              |-  0.25-  0.30
 |                             |                              |-  0.20-  0.25
 |                             |                              |-  0.15-  0.20
 |                             |                              |-  0.10-  0.15
 |                             |           4                  |-  0.05-  0.10
 |------------------------------------------------------------|-  0.00-  0.05
 |                             |                              |- -0.05-  0.00
 |                             |                              |- -0.10- -0.05
 |                             |                              |- -0.15- -0.10
 |                             |                              |- -0.20- -0.15
 |   5                         |                              |- -0.25- -0.20
 |                             |                              |- -0.30- -0.25
 |  6                          |                              |- -0.35- -0.30
 |                             |                              |- -0.40- -0.35
 |                             |                              |- -0.45- -0.40
 |                             |                              |- -0.50- -0.45
 |                             |                              |- -0.55- -0.50
 |                             |                              |- -0.60- -0.55
 |                             |                              |- -0.65- -0.60
 |                             |                              |- -0.70- -0.65
 |                             |                              |- -0.75- -0.70
 |                             |                              |- -0.80- -0.75
 |                             |                              |- -0.85- -0.80
 |                             |                              |- -0.90- -0.85
 |                             |                              |- -0.95- -0.90
 |                             |                              |- -1.00- -0.95
---------------------------------------------------------------
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Factor 1
-1.0-0.9-0.7-0.6-0.5-0.3-0.2-0.1 0.1 0.2 0.3 0.5 0.6 0.7 0.9 1.0
```

```
Labels:
 1 = situps
 2 = jumps
 3 = chins
 4 = pulse
 5 = weight
 6 = waist

SUBJECT FACTOR SCORE RESULTS:

Regression Coefficients with   20 valid cases.


Variables
               Factor 1     Factor 2
   weight       -0.418        0.150
    waist       -0.608        0.080
    pulse        0.042       -0.020
    chins       -0.024        0.203
   situps       -0.069        0.526
     jumps      -0.163        0.399


Standard Error of Factor Scores:
Factor 1    0.946
Factor 2    0.905
```

We note that two factors were extracted with eigenvalues greater than 1.0 and when rotated indicate that the three body measurements appear to load on one factor and that the performance measures load on the second factor. The data grid also now contains the "least-squares" factor scores for each subject. Hummm! I wonder what a hierarchical grouping of these subjects on the two factor scores would produce!

Correspondence Analysis

Correspondence analysis is a method for examining the relationship between two sets of categorical variables much as in a Chi-Squared analysis of a two-way contingency table. In fact, a typical chi-squared analysis is completed as part of this procedure. In addition, visualization of the relationships among the columns or rows of the analysis is performed in a manner similar to factor analysis. The data analyzed in the visualization is the table of relative proportions, that is, the original frequency values divided by the sum of all frequencies. The relative proportions of the row sums and the column sums are termed the "masses" of the rows or columns.

The method used to analyze the relative proportions involves what is now called the "Generalized Singular Value Decomposition" or more simply the generalized SVD. This method obtains roots and vectors of a rectangular matrix by decomposing that matrix into three portions: a matrix of left singular column vectors (A) that has n rows and q columns ($n \geq q$), a square diagonal matrix with q rows and columns of singular values (D), and a transposed matrix (B') that is m x q in size of right generalized singular vectors ($m = q-1$). Completing this analysis involves several steps. The first is to obtain the (regular) SVD analysis of a matrix Q defined as $D_r^{-1/2} P D_c^{-1/2}$ where D_r and D_c are diagonal matrices of row and column relative proportions and P is the matrix of relative proportions. The SVD of Q gives

$Q = U D V'$ where D is the desired diagonal matrix of eigenvalues and $U'U = V'V = I$. It should be noted that the first of the q roots is trivial and to be ignored. At this point we obtain $A = D_r^{1/2}$ Uand $B = D_c^{1/2} V$. The results of this SVD analysis is available on the output. Now $P = ADB'$. The row coordinates F and column coordinates G are then computed according to the table below:

Analysis Choice	Button Selected	Row Coordinates	Column Coordinates
Row Profile	Row	$F = D_r^{-1}AD$	$G = D_c^{-1}B$
Column Profile	Column	$F = D_r^{-1}A$	$G = D_c^{-1}BD$
Both Profiles	Both	$F = D_r^{-1}AD$	$G = D_c^{-1}BD$

If Row profiles are computed, the row coordinates are weighted centroids of the column coordinates and the inertias D^2 refer only to the row points. If the column profiles are computed, the column coordinates are weighted eentroids of the row coordinates and the inertias D^2 refer only to the column points. If both profiles are selected, neither row or column coordinates are weighted centroids of the other but the inertias D^2 refer to both sets of points. The q-1 inertias are plotted in a manner similar to a scree plot of roots in a factor analysis. The total inertia is, in fact, the chi-squared statistic divided by the total of all cell frequencies.

You may elect to plot the coordinates for any two pairs of coordinates. This will provide a graphical representation of the separation of the row or column categories similar to a plot of variables in a discriminant function analysis or factors in a factor analysis. A way of looking at correspondence analysis is to consider it as a method for decomposing the overall inertia by identifying a small number of dimensions in which the deviations from the expected values can be represented. This is similar to factor analysis where the total variance is decomposed so as to arrive at a lower dimensional representation of variables.

An example is the file labeled "Smokers.LAZ" that we will use for a correspondence analysis. The specifications form is shown below:

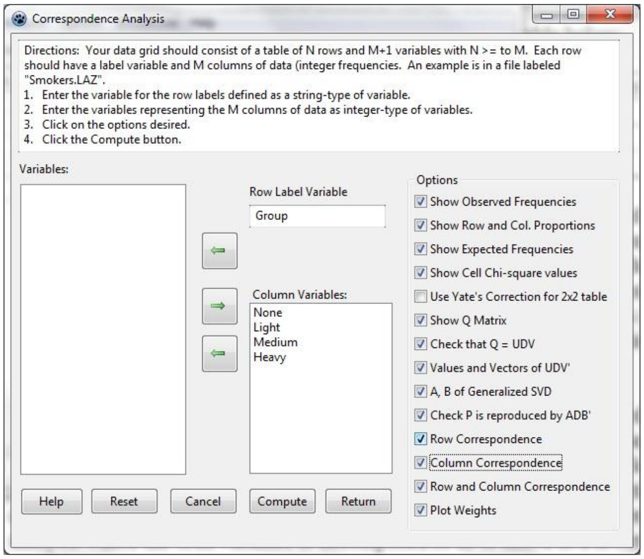


Fig. 7.12 Correspondence Analysis Form

When you click the "Compute" button you obtain the following:

CORRESPONDENCE ANALYSIS
Based on formulations of Bee-Leng Lee
Chapter 11 Correspondence Analysis for ViSta
Results are based on the Generalized Singular Value Decomposition
of P := A x D x B where P is the relative frequencies observed,
A are the left generalized singular vectors,
D is a diagonal matrix of generalized singular values, and
B is the transpose of the right generalized singular vectors.
NOTE: The first value and corresponding vectors are 1 and are
to be ignored.
An intermediate step is the regular SVD of the matrix Q := UDV
where Q := Dr^-1/2 x P x Dc^-1/2 where Dr is a diagonal matrix
of total row relative frequencies and Dc is a diagonal matrix
of total column relative frequencies.
Chi-square Analysis Results
No. of Cases := 193

OBSERVED FREQUENCIES

Frequencies

Variables	None	Light	Medium	Heavy	Total
Senior_Mgr.	4	2	3	2	11
Junior_Mgr.	4	3	7	4	18
Senior_Emp.	25	10	12	4	51
Junior_Emp.	18	24	33	13	88
Secretaries	10	6	7	2	25
Total	61	45	62	25	193

EXPECTED FREQUENCIES with 5 cases.

Variables	None	Light	Medium	Heavy
Senior_Mgr.	3.477	2.565	3.534	1.425
Junior_Mgr.	5.689	4.197	5.782	2.332
Senior_Emp.	16.119	11.891	16.383	6.606
Junior_Emp.	27.813	20.518	28.269	11.399
Secretaries	7.902	5.829	8.031	3.238

ROW PROPORTIONS with 5 cases.

```
Variables
                     None        Light       Medium        Heavy        Total
Senior_Mgr.         0.364        0.182        0.273        0.182        1.000
Junior_Mgr.         0.222        0.167        0.389        0.222        1.000
Senior_Emp.         0.490        0.196        0.235        0.078        1.000
Junior_Emp.         0.205        0.273        0.375        0.148        1.000
Secretaries         0.400        0.240        0.280        0.080        1.000
     Total         0.316        0.233        0.321        0.130        1.000
COLUMN PROPORTIONS with    5 cases.


Variables
                     None        Light       Medium        Heavy        Total
Senior_Mgr.         0.066        0.044        0.048        0.080        0.057
Junior_Mgr.         0.066        0.067        0.113        0.160        0.093
Senior_Emp.         0.410        0.222        0.194        0.160        0.264
Junior_Emp.         0.295        0.533        0.532        0.520        0.456
Secretaries         0.164        0.133        0.113        0.080        0.130
     Total         1.000        1.000        1.000        1.000        1.000
PROPORTIONS OF TOTAL N with    5 cases.


Variables
                     None        Light       Medium        Heavy        Total
Senior_Mgr.         0.021        0.010        0.016        0.010        0.057
Junior_Mgr.         0.021        0.016        0.036        0.021        0.093
Senior_Emp.         0.130        0.052        0.062        0.021        0.264
Junior_Emp.         0.093        0.124        0.171        0.067        0.456
Secretaries         0.052        0.031        0.036        0.010        0.130
     Total         0.316        0.233        0.321        0.130        1.000
CHI-SQUARED VALUE FOR CELLS with    5 cases.


Variables
                     None        Light       Medium        Heavy
Senior_Mgr.         0.079        0.124        0.081        0.232
Junior_Mgr.         0.502        0.341        0.256        1.194
Senior_Emp.         4.893        0.301        1.173        1.028
Junior_Emp.         3.463        0.591        0.792        0.225
Secretaries         0.557        0.005        0.132        0.474
Chi-square :=   16.442 with D.F. := 12. Prob. > value :=    0.172

Liklihood Ratio :=   16.348 with prob. > value := 0.1758

phi correlation := 0.2919

Pearson Correlation r := 0.0005

Mantel-Haenszel Test of Linear Association :=    0.000 with probability > value := 0.9999

The coefficient of contingency :=    0.280

Cramers V :=    0.169
Q Matrix with    5 cases.


Variables
                     None        Light       Medium        Heavy
Senior_Mgr.         0.154        0.090        0.115        0.121
Junior_Mgr.         0.121        0.105        0.210        0.189
Senior_Emp.         0.448        0.209        0.213        0.112
Junior_Emp.         0.246        0.381        0.447        0.277
Secretaries         0.256        0.179        0.178        0.080
P =  with    5 cases.


Variables
                     None        Light       Medium        Heavy
Senior_Mgr.         0.012       -0.002        0.041        0.003
Junior_Mgr.        -2.138       -0.463       -0.305        0.015
Senior_Emp.         3.911        0.778        0.679       -0.066
Junior_Emp.        -1.342       -0.451       -0.651        0.077
Secretaries         0.000        0.000        0.000        0.000
Inertia :=   0.0852
```

```
Row Dimensions (Ignore Col. 1 with    5 cases.


Variables
                  None        Light       Medium        Heavy
Senior_Mgr.       0.396       -0.267       13.267        8.441
Junior_Mgr.     -43.680      -39.212      -60.876       21.692
Senior_Emp.      28.202       23.251       47.807      -34.659
Junior_Emp.      -5.610       -7.813      -26.547       23.555
Secretaries       0.000        0.000        0.000        0.000
```



```
Row Dimensions (Ignore Col. 1 with    5 cases.


Variables
                  None        Light       Medium        Heavy
Senior_Mgr.       0.396       -0.267       13.267        8.441
Junior_Mgr.     -43.680      -39.212      -60.876       21.692
Senior_Emp.      28.202       23.251       47.807      -34.659
Junior_Emp.      -5.610       -7.813      -26.547       23.555
Secretaries       0.000        0.000        0.000        0.000
```



```
Column Dimensions (Ignore Col. 1) with    5 cases.


Variables
                  None        Light       Medium        Heavy
      None      -0.893        1.382       -0.219       -0.044
     Light      -0.261       -0.101        0.372        0.279
    Medium      -0.097       -0.066        0.004       -0.120
     Heavy      -0.021       -0.019       -0.040        0.026
```

Bartlett's Test of Sphericity

In matrix algebra, the determinate of an identity matrix is equal to 1.0.

The procedure calculates the determinate of the matrix of the sums of products and cross-products (S) from which an intercorrelation matrix is derived.

The determinant of the matrix S is converted to a chi-square statistic and tested for significance.

The null hypothesis is that the intercorrelation matrix comes from a population in which the variables are noncollinear (i.e. an identity matrix) and that the non-zero correlations in the sample matrix are due to sampling error.

Statistical Decision: if the sample intercorrelation matrix did not come from a population in which the intercorrelation matrix is an identity matrix the probability of the chi-square value will be small.

The example below uses the “Cansas.LAZ” file:

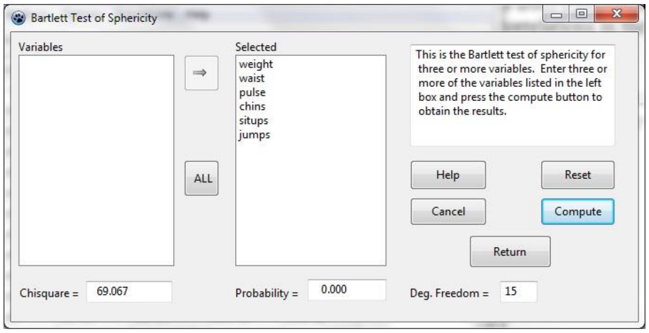


Fig. 7.13 The Bartlett Test of Sphericity Form

The results obtained are:

```
CORRELATION MATRIX with   20 cases.


Variables
                 weight        waist        pulse        chins       situps
    weight       1.000         0.870       -0.366       -0.390       -0.493
     waist       0.870         1.000       -0.353       -0.552       -0.646
     pulse      -0.366        -0.353        1.000        0.151        0.225
     chins      -0.390        -0.552        0.151        1.000        0.696
    situps      -0.493        -0.646        0.225        0.696        1.000
     jumps      -0.226        -0.191        0.035        0.496        0.669


Variables
                  jumps
    weight      -0.226
     waist      -0.191
     pulse       0.035
     chins       0.496
    situps       0.669
     jumps       1.000

Determinant of matrix =    0.021
chisquare =   69.067, D.F. = 15, Proabability greater value =    0.000
```

Log Linear Analysis for Cross-Classified Data

The contingency chi-square test for independence of two categorical variables is often employed in elementary statistical applications. However, the cell frequencies in a table can also be modeled as a regression model where the frequency is a function of the weighted sum of row effects, column effects, interaction effects and error. In practice the log of the frequencies is the dependent variable. This linear model can be expanded to three way or more tables. The investigator will often want to test the individual effects of rows, columns, slices, two-way interactions, three-way interactions, etc.

Three procedures are available for log linear analysis of classification data. These are described and illustrated in each of the sections below.

Log Linear for an A x B Classification Table I
Log Linear Analysis for an A x B x C Classification Table
Log Linear Screen

Log Linear for an A x B Classification Table

When you elect this analysis you see the dialogue boxes shown below. The difference depends on whether you are entering data from the main grid or if you are entering data directly on the form. In our example, we are entering data stored in a file labeled "ABCLogLinData.LAZ" and loaded into the Main Form grid. The results of the analysis is shown below these dialogue boxes. Each parameter is tested using the "G" statistic which is approximately chi-squared.

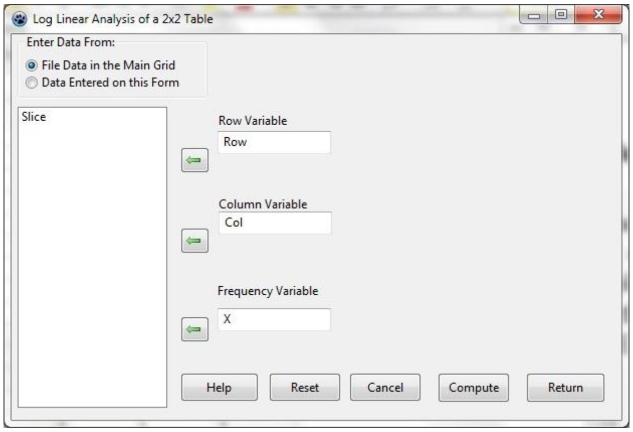


Fig. 7.14   AxB Log Linear Analysis Dialogue Form

```
ANALYSES FOR AN I BY J CLASSIFICATION TABLE

Reference: G.J.G. Upton, The Analysis of Cross-tabulated Data, 1980

Cross-Products Odds Ratio =  1.583
Log odds of the cross-products ratio =  0.460

Saturated Model Results

Observed Frequencies
ROW/COL        1         2      TOTAL
     1        27.00     36.00     63.00
     2        27.00     57.00     84.00
TOTAL         54.00     93.00    147.00

Log frequencies, row average and column average of log frequencies
ROW/COL        1         2      TOTAL
     1         3.30      3.58      3.44
     2         3.30      4.04      3.67
TOTAL          3.30      3.81      3.55

Expected Frequencies
ROW/COL        1         2      TOTAL
     1        27.00     36.00     63.00
     2        27.00     57.00     84.00
TOTAL         54.00     93.00    147.00

Cell Parameters
ROW COL   MU      LAMBDA ROW   LAMBDA COL   LAMBDA ROW x COL
  1   1  3.555     -0.115       -0.259         0.115
  1   2  3.555     -0.115        0.259        -0.115
  2   1  3.555      0.115       -0.259        -0.115
  2   2  3.555      0.115        0.259         0.115

Y squared statistic for model fit = -0.000 D.F. = 0
Independent Effects Model Results

Expected Frequencies
```

```
ROW/COL        1         2      TOTAL
     1       23.14     39.86     63.00
     2       30.86     53.14     84.00
TOTAL        54.00     93.00    147.00

Cell Parameters
ROW COL   MU      LAMBDA ROW   LAMBDA COL   LAMBDA ROW x COL
  1   1  3.557     -0.144       -0.272        0.000
  1   2  3.557     -0.144        0.272        0.000
  2   1  3.557      0.144       -0.272        0.000
  2   2  3.557      0.144        0.272        0.000


Y squared statistic for model fit =  1.773 D.F. = 1
Chi-squared =  1.778 with 1 D.F.
No Column Effects Model Results

Expected Frequencies
ROW/COL        1         2      TOTAL
     1       31.50     31.50     63.00
     2       42.00     42.00     84.00
TOTAL        73.50     73.50    147.00

Cell Parameters
ROW COL   MU      LAMBDA ROW   LAMBDA COL   LAMBDA ROW x COL
  1   1  3.594     -0.144        0.000       -0.000
  1   2  3.594     -0.144        0.000       -0.000
  2   1  3.594      0.144        0.000       -0.000
  2   2  3.594      0.144        0.000       -0.000


Y squared statistic for model fit = 12.245 D.F. = 2
No Row Effects Model Results

Expected Frequencies
ROW/COL        1         2      TOTAL
     1       27.00     46.50     73.50
     2       27.00     46.50     73.50
TOTAL        54.00     93.00    147.00

Cell Parameters
ROW COL   MU      LAMBDA ROW   LAMBDA COL   LAMBDA ROW x COL
  1   1  3.568      0.000       -0.272        0.000
  1   2  3.568      0.000        0.272        0.000
  2   1  3.568      0.000       -0.272        0.000
  2   2  3.568      0.000        0.272        0.000


Y squared statistic for model fit =  4.783 D.F. = 2
Equiprobability Effects Model Results

Expected Frequencies
ROW/COL        1         2      TOTAL
     1       36.75     36.75     36.75
     2       36.75     36.75     36.75
TOTAL        36.75     36.75    147.00

Cell Parameters
ROW COL   MU      LAMBDA ROW   LAMBDA COL   LAMBDA ROW x COL
  1   1  3.604      0.000        0.000        0.000
  1   2  3.604      0.000        0.000        0.000
  2   1  3.604      0.000        0.000        0.000
  2   2  3.604      0.000        0.000        0.000


Y squared statistic for model fit = 15.255 D.F. = 3
```

Log Linear Analysis for an A x B x C Classification Table

The three-way classification table can result in a number of linear models to describe the log of the observed frequencies as a function of row, column, slice, two-way interactions and the three-way interaction. When you select this option you see the dialogue box shown below. Notice that the option is given for entering data directly in the box if preferred.

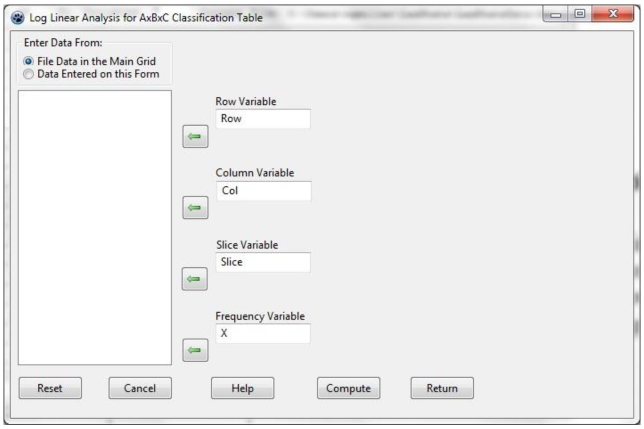


Fig. 7.15 AxBxC Classification Log Linear Dialogue

The following is the result of an analysis of three categorical variables stored in a file labeled "ABCLogLinData.LAZ".

```
Log-Linear Analysis of a Three Dimension Table

Observed Frequencies
  1   1   1      6.000
  1   1   2      9.000
  1   1   3     12.000
  1   2   1     15.000
  1   2   2     12.000
  1   2   3      9.000
  2   1   1      6.000
  2   1   2     15.000
  2   1   3      6.000
  2   2   1     15.000
  2   2   2     18.000
  2   2   3     24.000
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Sub-matrix AB
ROW/COL     1         2
    1     27.000    36.000
    2     27.000    57.000

Sub-matrix AC

ROW/COL     1         2         3
    1     21.000    21.000    21.000
```

```
    2     21.000    33.000    30.000

Sub-matrix BC

ROW/COL     1         2         3
    1     12.000    24.000    18.000
    2     30.000    30.000    33.000
Saturated Model

Expected Frequencies
  1   1   1      6.000
  1   1   2      9.000
  1   1   3     12.000
  1   2   1     15.000
  1   2   2     12.000
  1   2   3      9.000
  2   1   1      6.000
  2   1   2     15.000
  2   1   3      6.000
  2   2   1     15.000
  2   2   2     18.000
  2   2   3     24.000
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      1.792
  1   1   2      2.197
  1   1   3      2.485
  1   2   1      2.708
  1   2   2      2.485
  1   2   3      2.197
  2   1   1      1.792
  2   1   2      2.708
  2   1   3      1.792
  2   2   1      2.708
  2   2   2      2.890
  2   2   3      3.178
Totals for Dimension A
Row 1    2.311
Row 2    2.511
Totals for Dimension B
Col 1    2.128
Col 2    2.694
Totals for Dimension C
Slice 1    2.250
Slice 2    2.570
Slice 3    2.413
Cell Parameters
ROW COL SLICE     MU         LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.411       -0.100       -0.283       -0.161
                0.131        0.100       -0.175       -0.131

  1   1   2     2.411       -0.100       -0.283        0.159
                0.131       -0.129        0.166       -0.157

  1   1   3     2.411       -0.100       -0.283        0.002
                0.131        0.028        0.009        0.288

  1   2   1     2.411       -0.100        0.283       -0.161
               -0.131        0.100        0.175        0.131

  1   2   2     2.411       -0.100        0.283        0.159
               -0.131       -0.129       -0.166        0.157
```

```
  1   2   3     2.411       -0.100        0.283        0.002
               -0.131        0.028       -0.009       -0.288

  2   1   1     2.411        0.100       -0.283       -0.161
               -0.131       -0.100       -0.175        0.131

  2   1   2     2.411        0.100       -0.283        0.159
               -0.131        0.129        0.166        0.157

  2   1   3     2.411        0.100       -0.283        0.002
               -0.131       -0.028        0.009       -0.288

  2   2   1     2.411        0.100        0.283       -0.161
                0.131       -0.100        0.175       -0.131

  2   2   2     2.411        0.100        0.283        0.159
                0.131        0.129       -0.166       -0.157

  2   2   3     2.411        0.100        0.283        0.002
                0.131       -0.028       -0.009        0.288


G squared statistic for model fit =  0.000 D.F. = 0
Model of Independence

Expected Frequencies
  1   1   1      6.612
  1   1   2      8.501
  1   1   3      8.029
  1   2   1     11.388
  1   2   2     14.641
  1   2   3     13.828
  2   1   1      8.816
  2   1   2     11.335
  2   1   3     10.706
  2   2   1     15.184
  2   2   2     19.522
  2   2   3     18.437
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      1.889
  1   1   2      2.140
  1   1   3      2.083
  1   2   1      2.433
  1   2   2      2.684
  1   2   3      2.627
  2   1   1      2.177
  2   1   2      2.428
  2   1   3      2.371
  2   2   1      2.720
  2   2   2      2.972
  2   2   3      2.914
Totals for Dimension A
Row 1    2.309
Row 2    2.597
Totals for Dimension B
Col 1    2.181
Col 2    2.725
Totals for Dimension C
Slice 1    2.305
Slice 2    2.556
Slice 3    2.499
Cell Parameters
```

```
ROW COL SLICE      MU         LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1      2.453       -0.144       -0.272       -0.148
                 0.000        0.000        0.000       -0.000

  1   1   2      2.453       -0.144       -0.272        0.103
                 0.000       -0.000        0.000        0.000

  1   1   3      2.453       -0.144       -0.272        0.046
                 0.000        0.000        0.000        0.000

  1   2   1      2.453       -0.144        0.272       -0.148
                 0.000        0.000        0.000        0.000

  1   2   2      2.453       -0.144        0.272        0.103
                 0.000       -0.000       -0.000        0.000

  1   2   3      2.453       -0.144        0.272        0.046
                 0.000        0.000       -0.000        0.000

  2   1   1      2.453        0.144       -0.272       -0.148
                 0.000        0.000        0.000       -0.000

  2   1   2      2.453        0.144       -0.272        0.103
                 0.000       -0.000        0.000        0.000

  2   1   3      2.453        0.144       -0.272        0.046
                 0.000        0.000        0.000       -0.000

  2   2   1      2.453        0.144        0.272       -0.148
                -0.000        0.000        0.000        0.000

  2   2   2      2.453        0.144        0.272        0.103
                -0.000       -0.000       -0.000        0.000

  2   2   3      2.453        0.144        0.272        0.046
                -0.000        0.000       -0.000        0.000

G squared statistic for model fit = 11.471 D.F. = 7
No AB Effect

Expected Frequencies
  1   1   1      6.000
  1   1   2      9.333
  1   1   3      7.412
  1   2   1     15.000
  1   2   2     11.667
  1   2   3     13.588
  2   1   1      6.000
  2   1   2     14.667
  2   1   3     10.588
  2   2   1     15.000
  2   2   2     18.333
  2   2   3     19.412
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      1.792
  1   1   2      2.234
  1   1   3      2.003
  1   2   1      2.708
  1   2   2      2.457
  1   2   3      2.609
```

```
  2   1   1      1.792
  2   1   2      2.686
  2   1   3      2.360
  2   2   1      2.708
  2   2   2      2.909
  2   2   3      2.966
Totals for Dimension A
Row 1    2.300
Row 2    2.570
Totals for Dimension B
Col 1    2.144
Col 2    2.726
Totals for Dimension C
Slice 1    2.250
Slice 2    2.571
Slice 3    2.484
Cell Parameters
ROW COL SLICE     MU        LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.435       -0.135       -0.291       -0.185
                0.000        0.135       -0.167        0.000

  1   1   2     2.435       -0.135       -0.291        0.136
                0.000       -0.091        0.179        0.000

  1   1   3     2.435       -0.135       -0.291        0.049
                0.000       -0.044       -0.012        0.000

  1   2   1     2.435       -0.135        0.291       -0.185
                0.000        0.135        0.167        0.000

  1   2   2     2.435       -0.135        0.291        0.136
                0.000       -0.091       -0.179        0.000

  1   2   3     2.435       -0.135        0.291        0.049
                0.000       -0.044        0.012        0.000

  2   1   1     2.435        0.135       -0.291       -0.185
                0.000       -0.135       -0.167       -0.000

  2   1   2     2.435        0.135       -0.291        0.136
                0.000        0.091        0.179       -0.000

  2   1   3     2.435        0.135       -0.291        0.049
                0.000        0.044       -0.012       -0.000

  2   2   1     2.435        0.135        0.291       -0.185
                0.000       -0.135        0.167        0.000

  2   2   2     2.435        0.135        0.291        0.136
                0.000        0.091       -0.179        0.000

  2   2   3     2.435        0.135        0.291        0.049
                0.000        0.044        0.012        0.000


G squared statistic for model fit =  7.552 D.F. = 3
No AC Effect

Expected Frequencies
  1   1   1      6.000
  1   1   2     12.000
  1   1   3      9.000
  1   2   1     11.613
  1   2   2     11.613
  1   2   3     12.774
  2   1   1      6.000
  2   1   2     12.000
  2   1   3      9.000
  2   2   1     18.387
  2   2   2     18.387
  2   2   3     20.226
Totals for Dimension A
```

```
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      1.792
  1   1   2      2.485
  1   1   3      2.197
  1   2   1      2.452
  1   2   2      2.452
  1   2   3      2.547
  2   1   1      1.792
  2   1   2      2.485
  2   1   3      2.197
  2   2   1      2.912
  2   2   2      2.912
  2   2   3      3.007
Totals for Dimension A
Row 1    2.321
Row 2    2.551
Totals for Dimension B
Col 1    2.158
Col 2    2.714
Totals for Dimension C
Slice 1    2.237
Slice 2    2.583
Slice 3    2.487
Cell Parameters
ROW COL SLICE     MU        LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.436       -0.115       -0.278       -0.199
                0.115        0.000       -0.167        0.000

  1   1   2     2.436       -0.115       -0.278        0.148
                0.115        0.000        0.179        0.000

  1   1   3     2.436       -0.115       -0.278        0.051
                0.115       -0.000       -0.012        0.000

  1   2   1     2.436       -0.115        0.278       -0.199
               -0.115        0.000        0.167        0.000

  1   2   2     2.436       -0.115        0.278        0.148
               -0.115        0.000       -0.179        0.000

  1   2   3     2.436       -0.115        0.278        0.051
               -0.115       -0.000        0.012        0.000

  2   1   1     2.436        0.115       -0.278       -0.199
               -0.115        0.000       -0.167       -0.000

  2   1   2     2.436        0.115       -0.278        0.148
               -0.115        0.000        0.179       -0.000

  2   1   3     2.436        0.115       -0.278        0.051
               -0.115        0.000       -0.012       -0.000

  2   2   1     2.436        0.115        0.278       -0.199
                0.115        0.000        0.167       -0.000

  2   2   2     2.436        0.115        0.278        0.148
                0.115        0.000       -0.179       -0.000

  2   2   3     2.436        0.115        0.278        0.051
                0.115        0.000        0.012       -0.000
```

```
G squared statistic for model fit =  7.055 D.F. = 4
No BC Effect

Expected Frequencies
  1   1   1      9.000
  1   1   2      9.000
  1   1   3      9.000
  1   2   1     12.000
  1   2   2     12.000
  1   2   3     12.000
  2   1   1      6.750
  2   1   2     10.607
  2   1   3      9.643
  2   2   1     14.250
  2   2   2     22.393
  2   2   3     20.357
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      2.197
  1   1   2      2.197
  1   1   3      2.197
  1   2   1      2.485
  1   2   2      2.485
  1   2   3      2.485
  2   1   1      1.910
  2   1   2      2.362
  2   1   3      2.266
  2   2   1      2.657
  2   2   2      3.109
  2   2   3      3.013
Totals for Dimension A
Row 1    2.341
Row 2    2.553
Totals for Dimension B
Col 1    2.188
Col 2    2.706
Totals for Dimension C
Slice 1    2.312
Slice 2    2.538
Slice 3    2.490
Cell Parameters
ROW COL SLICE     MU        LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.447       -0.106       -0.259       -0.135
                0.115        0.135        0.000       -0.000

  1   1   2     2.447       -0.106       -0.259        0.091
                0.115       -0.091        0.000       -0.000

  1   1   3     2.447       -0.106       -0.259        0.044
                0.115       -0.044       -0.000        0.000

  1   2   1     2.447       -0.106        0.259       -0.135
               -0.115        0.135       -0.000        0.000

  1   2   2     2.447       -0.106        0.259        0.091
               -0.115       -0.091       -0.000        0.000

  1   2   3     2.447       -0.106        0.259        0.044
               -0.115       -0.044       -0.000        0.000

  2   1   1     2.447        0.106       -0.259       -0.135
               -0.115       -0.135        0.000        0.000
```

```
  2   1   2     2.447        0.106       -0.259        0.091
               -0.115        0.091        0.000        0.000

  2   1   3     2.447        0.106       -0.259        0.044
               -0.115        0.044       -0.000        0.000

  2   2   1     2.447        0.106        0.259       -0.135
                0.115       -0.135       -0.000        0.000

  2   2   2     2.447        0.106        0.259        0.091
                0.115        0.091       -0.000        0.000

  2   2   3     2.447        0.106        0.259        0.044
                0.115        0.044       -0.000        0.000


G squared statistic for model fit =  8.423 D.F. = 4
Model of No Slice (C) effect

Expected Frequencies
  1   1   1      7.714
  1   1   2      7.714
  1   1   3      7.714
  1   2   1     13.286
  1   2   2     13.286
  1   2   3     13.286
  2   1   1     10.286
  2   1   2     10.286
  2   1   3     10.286
  2   2   1     17.714
  2   2   2     17.714
  2   2   3     17.714
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
Slice 1   49.000
Slice 2   49.000
Slice 3   49.000

Log Frequencies
  1   1   1      2.043
  1   1   2      2.043
  1   1   3      2.043
  1   2   1      2.587
  1   2   2      2.587
  1   2   3      2.587
  2   1   1      2.331
  2   1   2      2.331
  2   1   3      2.331
  2   2   1      2.874
  2   2   2      2.874
  2   2   3      2.874
Totals for Dimension A
Row 1    2.315
Row 2    2.603
Totals for Dimension B
Col 1    2.187
Col 2    2.731
Totals for Dimension C
Slice 1    2.459
Slice 2    2.459
Slice 3    2.459
Cell Parameters
ROW COL SLICE     MU        LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.459       -0.144       -0.272        0.000
                0.000        0.000        0.000       -0.000
```

```
  1   1   2     2.459       -0.144       -0.272        0.000
                0.000        0.000        0.000       -0.000

  1   1   3     2.459       -0.144       -0.272        0.000
                0.000        0.000        0.000       -0.000

  1   2   1     2.459       -0.144        0.272        0.000
                0.000        0.000        0.000        0.000

  1   2   2     2.459       -0.144        0.272        0.000
                0.000        0.000        0.000        0.000

  1   2   3     2.459       -0.144        0.272        0.000
                0.000        0.000        0.000        0.000

  2   1   1     2.459        0.144       -0.272        0.000
                0.000        0.000        0.000       -0.000

  2   1   2     2.459        0.144       -0.272        0.000
                0.000        0.000        0.000       -0.000

  2   1   3     2.459        0.144       -0.272        0.000
                0.000        0.000        0.000       -0.000

  2   2   1     2.459        0.144        0.272        0.000
               -0.000        0.000        0.000        0.000

  2   2   2     2.459        0.144        0.272        0.000
               -0.000        0.000        0.000        0.000

  2   2   3     2.459        0.144        0.272        0.000
               -0.000        0.000        0.000        0.000


G squared statistic for model fit = 13.097 D.F. = 9
Model of no Column (B) effect

Expected Frequencies
  1   1   1      9.000
  1   1   2     11.571
  1   1   3     10.929
  1   2   1      9.000
  1   2   2     11.571
  1   2   3     10.929
  2   1   1     12.000
  2   1   2     15.429
  2   1   3     14.571
  2   2   1     12.000
  2   2   2     15.429
  2   2   3     14.571
Totals for Dimension A
Row 1   63.000
Row 2   84.000
Totals for Dimension B
Col 1   73.500
Col 2   73.500
Totals for Dimension C
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      2.197
  1   1   2      2.449
  1   1   3      2.391
  1   2   1      2.197
  1   2   2      2.449
  1   2   3      2.391
  2   1   1      2.485
  2   1   2      2.736
  2   1   3      2.679
  2   2   1      2.485
  2   2   2      2.736
  2   2   3      2.679
```

```
Totals for Dimension A
Row 1    2.346
Row 2    2.633
Totals for Dimension B
Col 1    2.490
Col 2    2.490
Totals for Dimension C
Slice 1    2.341
Slice 2    2.592
Slice 3    2.535
Cell Parameters
ROW COL SLICE     MU        LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.490       -0.144       -0.000       -0.148
                0.000        0.000        0.000       -0.000

  1   1   2     2.490       -0.144       -0.000        0.103
                0.000        0.000        0.000       -0.000

  1   1   3     2.490       -0.144       -0.000        0.046
                0.000        0.000        0.000       -0.000

  1   2   1     2.490       -0.144       -0.000       -0.148
                0.000        0.000        0.000       -0.000

  1   2   2     2.490       -0.144       -0.000        0.103
                0.000        0.000        0.000       -0.000

  1   2   3     2.490       -0.144       -0.000        0.046
                0.000        0.000        0.000       -0.000

  2   1   1     2.490        0.144       -0.000       -0.148
                0.000        0.000        0.000       -0.000

  2   1   2     2.490        0.144       -0.000        0.103
                0.000        0.000        0.000       -0.000

  2   1   3     2.490        0.144       -0.000        0.046
                0.000        0.000        0.000       -0.000

  2   2   1     2.490        0.144       -0.000       -0.148
                0.000        0.000        0.000       -0.000

  2   2   2     2.490        0.144       -0.000        0.103
                0.000        0.000        0.000       -0.000

  2   2   3     2.490        0.144       -0.000        0.046
                0.000        0.000        0.000       -0.000


G squared statistic for model fit = 21.943 D.F. = 8
Model of no Row (A) effect

Expected Frequencies
  1   1   1      7.714
  1   1   2      9.918
  1   1   3      9.367
  1   2   1     13.286
  1   2   2     17.082
  1   2   3     16.133
  2   1   1      7.714
  2   1   2      9.918
  2   1   3      9.367
  2   2   1     13.286
  2   2   2     17.082
  2   2   3     16.133
Totals for Dimension A
Row 1   73.500
Row 2   73.500
Totals for Dimension B
Col 1   54.000
Col 2   93.000
Totals for Dimension C
```

```
Slice 1   42.000
Slice 2   54.000
Slice 3   51.000

Log Frequencies
  1   1   1      2.043
  1   1   2      2.294
  1   1   3      2.237
  1   2   1      2.587
  1   2   2      2.838
  1   2   3      2.781
  2   1   1      2.043
  2   1   2      2.294
  2   1   3      2.237
  2   2   1      2.587
  2   2   2      2.838
  2   2   3      2.781
Totals for Dimension A
Row 1    2.463
Row 2    2.463
Totals for Dimension B
Col 1    2.192
Col 2    2.735
Totals for Dimension C
Slice 1    2.315
Slice 2    2.566
Slice 3    2.509
Cell Parameters
ROW COL SLICE     MU         LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.463        0.000       -0.272       -0.148
                0.000       -0.000        0.000        0.000

  1   1   2     2.463        0.000       -0.272        0.103
                0.000       -0.000        0.000        0.000

  1   1   3     2.463        0.000       -0.272        0.046
                0.000       -0.000        0.000        0.000

  1   2   1     2.463        0.000        0.272       -0.148
               -0.000       -0.000        0.000        0.000

  1   2   2     2.463        0.000        0.272        0.103
               -0.000       -0.000        0.000        0.000

  1   2   3     2.463        0.000        0.272        0.046
               -0.000       -0.000        0.000        0.000

  2   1   1     2.463        0.000       -0.272       -0.148
                0.000       -0.000        0.000        0.000

  2   1   2     2.463        0.000       -0.272        0.103
                0.000       -0.000        0.000        0.000

  2   1   3     2.463        0.000       -0.272        0.046
                0.000       -0.000        0.000        0.000

  2   2   1     2.463        0.000        0.272       -0.148
               -0.000       -0.000        0.000        0.000

  2   2   2     2.463        0.000        0.272        0.103
               -0.000       -0.000        0.000        0.000

  2   2   3     2.463        0.000        0.272        0.046
               -0.000       -0.000        0.000        0.000


G squared statistic for model fit = 14.481 D.F. = 8
Equi-probability Model

Expected Frequencies
  1   1   1     12.250
  1   1   2     12.250
```

```
  1   1   3     12.250
  1   2   1     12.250
  1   2   2     12.250
  1   2   3     12.250
  2   1   1     12.250
  2   1   2     12.250
  2   1   3     12.250
  2   2   1     12.250
  2   2   2     12.250
  2   2   3     12.250
Totals for Dimension A
Row 1   73.500
Row 2   73.500
Totals for Dimension B
Col 1   73.500
Col 2   73.500
Totals for Dimension C
Slice 1   49.000
Slice 2   49.000
Slice 3   49.000

Log Frequencies
  1   1   1      2.506
  1   1   2      2.506
  1   1   3      2.506
  1   2   1      2.506
  1   2   2      2.506
  1   2   3      2.506
  2   1   1      2.506
  2   1   2      2.506
  2   1   3      2.506
  2   2   1      2.506
  2   2   2      2.506
  2   2   3      2.506
Totals for Dimension A
Row 1    2.506
Row 2    2.506
Totals for Dimension B
Col 1    2.506
Col 2    2.506
Totals for Dimension C
Slice 1    2.506
Slice 2    2.506
Slice 3    2.506
Cell Parameters
ROW COL SLICE     MU        LAMBDA A     LAMBDA B     LAMBDA C
               LAMBDA AB    LAMBDA AC    LAMBDA BC    LAMBDA ABC

  1   1   1     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  1   1   2     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  1   1   3     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  1   2   1     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  1   2   2     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  1   2   3     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  2   1   1     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  2   1   2     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000

  2   1   3     2.506        0.000        0.000        0.000
                0.000        0.000        0.000        0.000
```

```
  2   2   1      2.506        0.000        0.000        0.000
                 0.000        0.000        0.000        0.000

  2   2   2      2.506        0.000        0.000        0.000
                 0.000        0.000        0.000        0.000

  2   2   3      2.506        0.000        0.000        0.000
                 0.000        0.000        0.000        0.000

G squared statistic for model fit = 26.579 D.F. = 11
```

Log Linear Screen

A large number of possible parameters may be tested by the log linear procedures. It is not uncommon to complete an initial screening of the data for an analysis. In particular, an investigator may want to consider one of the variables as having "fixed" marginal values while the other margins are free to vary. These marginal associations can be tested by this procedure.

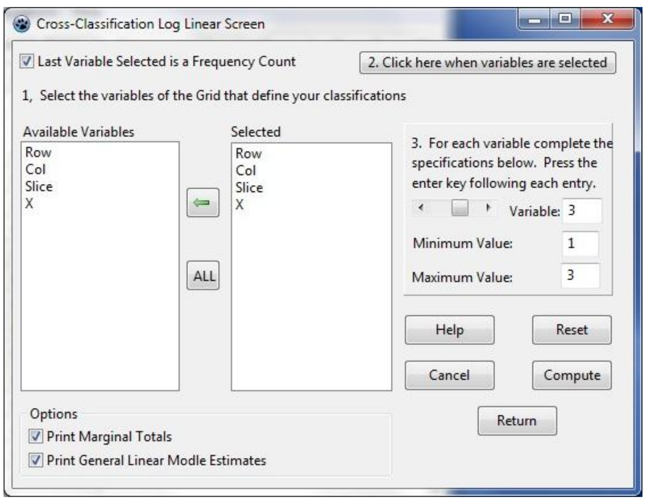


Fig. 7.15 Log Linear Screening Dialogue

Shown below are the results of the screen for a three-way classification table.

```
FILE: C:\lazarus\Projects\LazStats\LazStatsData\ABCLogLinData.LAZ


Marginal Totals for Row

       1        2
      63       84

Marginal Totals for Col

       1        2
      54       93

Marginal Totals for Slice

       1        2        3
      42       54       51

Total Frequencies = 147

EXPECTED CELL VALUES FOR MODEL OF COMPLETE INDEPENDENCE
```

```
Cell          Observed  Expected  Log Expected
  1   1   1          6      6.61      1.889
  2   1   1          6      8.82      2.177
  1   2   1         15     11.39      2.433
  2   2   1         15     15.18      2.720
  1   1   2          9      8.50      2.140
  2   1   2         15     11.34      2.428
  1   2   2         12     14.64      2.684
  2   2   2         18     19.52      2.972
  1   1   3         12      8.03      2.083
  2   1   3          6     10.71      2.371
  1   2   3          9     13.83      2.627
  2   2   3         24     18.44      2.914
Chisquare =     11.310 with probability =      0.004 (DF = 2)
G squared =     11.471 with probability =      0.003 (DF = 2)

U (mu) for general loglinear model =       2.45
First Order LogLinear Model Factors and N of Cells in Each
CELL               U1  N Cells    U2  N Cells    U3  N Cells

  1   1   1     -0.144   6     -0.272   6     -0.148   4
  2   1   1      0.144   6     -0.272   6     -0.148   4
  1   2   1     -0.144   6      0.272   6     -0.148   4
  2   2   1      0.144   6      0.272   6     -0.148   4
  1   1   2     -0.144   6     -0.272   6      0.103   4
  2   1   2      0.144   6     -0.272   6      0.103   4
  1   2   2     -0.144   6      0.272   6      0.103   4
  2   2   2      0.144   6      0.272   6      0.103   4
  1   1   3     -0.144   6     -0.272   6      0.046   4
  2   1   3      0.144   6     -0.272   6      0.046   4
  1   2   3     -0.144   6      0.272   6      0.046   4
  2   2   3      0.144   6      0.272   6      0.046   4
Second Order Loglinear Model Terms and N of Cells in Each
CELL              U12  N Cells  U13  N Cells  U23  N Cells

  1   1   1     -0.416   3    -0.292   2    -0.420   2
  2   1   1     -0.128   3    -0.005   2    -0.420   2
  1   2   1      0.128   3    -0.292   2     0.123   2
  2   2   1      0.416   3    -0.005   2     0.123   2
  1   1   2     -0.416   3    -0.041   2    -0.169   2
  2   1   2     -0.128   3     0.247   2    -0.169   2
  1   2   2      0.128   3    -0.041   2     0.375   2
  2   2   2      0.416   3     0.247   2     0.375   2
  1   1   3     -0.416   3    -0.098   2    -0.226   2
  2   1   3     -0.128   3     0.190   2    -0.226   2
  1   2   3      0.128   3    -0.098   2     0.317   2
  2   2   3      0.416   3     0.190   2     0.317   2
SCREEN FOR INTERACTIONS AMONG THE VARIABLES
Adapted from the Fortran program by Lustbader and Stodola printed in
Applied Statistics, Volume 30, Issue 1, 1981, pages 97-105 as Algorithm
AS 160 Partial and Marginal Association in Multidimensional Contingency Tables

Statistics for tests that the interactions of a given order are zero
ORDER     STATISTIC    D.F.         PROB.
    1         15.108        4       0.004
    2          6.143        5       0.293
    3          5.328        2       0.070

Statistics for Marginal Association Tests
VARIABLE  ASSOC.  PART ASSOC. MARGINAL ASSOC. D.F.    PROB
    1         1         3.010        3.010    1        0.083
    1         2        10.472       10.472    1        0.001
    1         3         1.626        1.626    2        0.444
    2         1         2.224        1.773    1        0.183
    2         2         1.726        1.275    2        0.529
    2         3         3.095        2.644    2        0.267
```

Chapter 8. Non-Parametric Statistics

Beginning statistics students are usually introduced to what are called "parametric" statistics methods. Those methods utilize "models" of score distributions such as the normal (Gaussian) distribution, Poisson distribution, binomial distribution, etc. The emphasis in parametric statistical methods is estimating population parameters from sample statistics when the distribution of the population scores can be assumed to be one of these theoretical models. The observations made are also assumed to be based on continous variables that utilize an interval or ratio scale of measurement. Frequently the measurement scales available yield only nominal or ordinal values and nothing can be assumed about the distribution of such values in the population sampled. If however, random sampling has been utilized in selecting subjects, one can still make inferences about relationships and differences similar to those made with parametric statistics. For example, if students enrolled in two courses are assigned a rank on their achievement in each of the two courses, it is reasonable to expect that students that rank high in one course would tend to rank high in the other course. Since a rank only indicates order however and not "how much" was achieved, we cannot use the usual product-moment correlation to indicate the relationship between the ranks. We can estimate, however, what the product of rank values in a group of n subjects where the ranks are randomly assigned would tend to be and estimate the variability of these sums or rank products for repeated samples. This would lead to a test of significance of the departure of our rank product sum (or average) from a value expected when there is no relationship.

A variety of non-parametric methods have been developed for nominal and ordinal measures to indicate congruence or similarity among independent groups or repeated measures on subjects in a group.

Contingency Chi-Square

The frequency chi-square statistic is used to accept or reject hypotheses concerning the degree to which observed frequencies depart from theoretical frequencies in a row by column contingency table with fixed marginal frequencies. It therefore tests the independence of the categorical variables defining the rows and columns. As an example, assume 50 males and 50 females are randomly assigned to each of three types of instructional methods to learn beginning French, (a) using a language laboratory, (b) using a computer with voice synthesizer and (c) using an advanced student tutor. Following a treatment period, a test is administered to each student with scoring results being pass or fail. The frequency of passing is then recorded for each cell in the 2 by 3 array (gender by treatment). If gender is independent of the treatment variable, the expected frequency of males that pass in each treatment would be the same as the expected frequency for females. The chi-squared statistic is obtained as

$$\chi^2 = \frac{\sum_{i=1}^{row} \sum_{j=1}^{col} (f_{ij} - F_{ij})^2}{F_{ij}} \qquad (8.1)$$

where f_{ij} is the observed frequency, F_{ij} the expected frequency, and χ^2 is the chi-squared statistic with degrees of freedom (rows - 1) times (columns - 1).

Spearman Rank Correlation

When the researcher's data represent ordinal measures such as ranks with some observations being tied for the same rank, the Rank Correlation may be the appropriate statistic to calculate. While the computation for the case of untied cases is the same as that for the Pearson Product-Moment correlation, the correction for tied ranks is found only in the Spearman correlation. In addition, the interpretation of the significance of the Rank Correlation may differ from that of the Pearson Correlation where bivariate normalcy is assumed.

Mann-Whitney U Test

An alternative to the Student t-test when the scale of measurement cannot be assumed to be interval or ratio and the distribution of errors is unknown is a non-parametric test known as the Mann-Whitney test. In this test, the dependent variable scores for both groups are ranked and the number of times that one groups scores exceed the rank of scores in the other group are recorded. This total number of times scores in one group exceed those of the other is named U. The sampling distribution of U is known and forms the basis for the hypothesis that the scores come from the same population.

Fisher's Exact Test

The probability of any given pattern of responses in a 2 by 2 table may be calculated from the hypergeometric probability distribution as

$$P = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!} \qquad (8.2)$$

where A, B, C, and D correspond to the frequencies in the four quadrants of the table and N corresponds to the total number of individuals sampled.

Kendall's Coefficient of Concordance

It is not uncommon that a group of people are asked to judge a group of persons or objects by rank ordering them from highest to lowest. It is then desirable to have some index of the degree to which the various judges agreed, that is, ranked the objects in the same order. The Coefficient of Concordance is a measure varying between 0 and 1 that indicates the degree of agreement among judges. It is defined as:

W = Variance of rank sums / maximum variance of rank sums.

The coefficient W may also be used to obtain the average rank correlation among the judges by the formula:

$$M_r = (mW - 1) / (m - 1) \qquad (8.3)$$

where M_r is the average (Spearman) rank correlation, m is the number of judges and W is the Coefficient of Concordance.

Kruskal-Wallis One-Way ANOVA

One-Way, Fixed-Effects Analysis of Variance assumes that error (residual) scores are normally distributed, that subjects are randomly selected from the population and assigned to treatments, and that the error scores are equally distributed in the populations representing the treatments. The scale of measurement for the dependent variable is assumed to be interval or ratio. But what can you do if, in fact, your measure is only ordinal (for example like most classroom tests), and you cannot assume normally distributed, homoscedastic error distributions?

Why, of course, you convert the scores to ranks and ask if the sum of rank scores in each treatment group are the same within sampling error! The Kruskal-Wallis One-Way Analysis of variance converts the dependent score for each subject in the study to a rank from 1 to N. It then examines the ranks attained by subjects in each of the treatment groups. Then a test statistic which is distributed as Chi-Squared with degrees of freedom equal to the number of treatment groups minus one is obtained from:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{K} R_j^2 / n_j - 3(N+1) \tag{8.4}$$

where N is the total number of subjects in the experiment, n_j is the number of subjects in the j<u>th</u> treatment, K is the number of treatments and R_j is the sum of ranks in the j<u>th</u> treatment.

Wilcoxon Matched-Pairs Signed Ranks Test

This test provides an alternative to the student t-test for matched score data where the assumptions for the parametric t-test cannot be met. In using this test, the difference is obtained between each of N pairs of scores observed on matched objects, for example, the difference between pretest and post-test scores for a group of students. The difference scores obtained are then ranked. The ranks of negative score differences are summed and the ranks of positive score differences are summed. The test statistic T is the smaller of these two sums. Difference scores of 0 are eliminated since a rank cannot be assigned. If the null hypothesis of no difference between the groups of scores is true, the sum of positive ranks should not differ from the sum of negative ranks beyond that expected by chance. Given N ranks, there is a finite number of ways of obtaining a given sum T. There are a total of 2 raised to the N ways of assigning positive and negative differences to N ranks. In a sample of 5 pairs, for example, there are 2 to the fifth power = 32 ways. Each rank sign would occur with probability of 1/32. The probability of getting a particular total T is

$$P_T = \frac{\text{Ways of getting T}}{2^N} \tag{8.5}$$

The cumulative probabilities for T, T-1,....,0 are obtained for the observed T value and reported. For large samples, a normally distributed z score is approximated and used.

Cochran Q Test

The Cochran Q test is used to test whether or not two or more matched sets of frequencies or proportions, measured on a nominal or ordinal scale, differ significantly among themselves. Typically, observations are dichotomous, that is, scored as 0 or 1 depending on whether or not the subject falls into one or the other criterion group. An example of research for which the Q test may be applied might be the agreement or disagreement to the question "Should abortions be legal?". The research design might call for a sample of n subjects answering the question prior to a debate and following a debate on the topic and subsequently six months later. The Q test applied to these data would test whether or not the proportion agreeing was the same under these three time periods. The Q statistic is obtained as

$$Q = \frac{(K-1)\sum_{j=1}^{K} G_j^2 - \left(\sum_{j=1}^{K} G_j\right)^2}{K\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2} \qquad (8.6)$$

where K is the number of treatments (groups of scores, G_j is the sum with the jth treatment group, and L_i is the sum within case i (across groups). The Q statistic is distributed approximately as Chi-squared with degrees of freedom K-1. If Q exceeds the Chi-Squared value corresponding to the cumulative probability value, the hypothesis of equal proportions for the K groups is rejected.

Sign Test

Imagine a counseling psychologist who sees, over a period of months, a number of clients with personal problems. Suppose the psychologist routinely contacts each client for a six month followup to see how they are doing. The counselor could make an estimate of client "adjustment" before treatment and at the followup time (or better still, have another person independently estimate adjustment at these two time periods). We may assume some underlying continuous "adjustment" variable even though we have no idea about the population distribution of the variable. We are intrested in knowing, of course, whether or not people are better adjusted six months after therapy than before. Note that we are only comparing the "before" and "after" state of the individuals with each other, not with other subjects. If we assign a + to the situation of improved adjustment and a - to the situation of same or poorer adjustment, we have the data required for a Sign Test. If treatment has had no effect, we would expect approximately one half the subjects would receive plus signs and the others negative signs. The sampling distribution of the proportion of plus signs is given by the binomial probability distribution with parameter of .5 and the number of events equal to n, the number of pairs of observations.

Friedman Two Way ANOVA

Imagine an experiment using, say, ten groups of subjects with four subjects in each group that have been matched on some relevant variables (or even using the same subjects). The matched subjects in each group are exposed to four different treatments such as teaching methods, dosages of medicine, proportion of positive responses to statements or questions, etc. Assume that some criterion measure on at least a nominal scale is available to measure the effect of each treatment. Now rank the subjects in each group on the basis of their scores on the criterion. We may now ask whether the ranks in each treatment come from the same population. Had we been able to assume an interval or ratio measure and normally distributed errors, we might have used a repeated measures analysis of variance. Failing to meet the parametric test assumptions, we instead examine the sum of ranks obtained under each of the treatment conditions and ask whether they differ significantly. The test statistic is distributed as Chi-squared with degrees of freedom equal to the number of treatments minus one. It is obtained as where N is the number of groups, K the number of treatments (or number of subjects in each group), and R_j is the sum of ranks in each treatment.

Probability of a Binomial Event

The BINOMIAL program is a short program to calculate the probability of obtaining k or fewer occurrences of a dichotomous variable out of a total of n observations when the probability of an occurrence is known. For example, assume a test consists of 5 multiple choice items with each item scored correct or incorrect. Also assume that there are five equally plausible choices for a student with no knowledge concerning any item. In this case, the probability of a student guessing the correct answer to a single item is 1/5 or .20 . We may use the binomial program to obtain the probabilities that a student guessing on each item of the test gets a score of 0, 1, 2, 3, 4, or 5 items correct by chance alone.

The formula for the probability of a dichotomous event k where the probability of a single event is p (and the probability of a non-event is q = 1 - p is given as:

$$P(k) = \frac{N!}{(N-k)!\,k!}\, p^{(N-k)} \quad q^{k} \tag{8.7}$$

For example, if a “fair” coin is tossed three times with the probabilities of heads is p = .5 (and q = .5) then the probabilty of observing 2 heads is

$$P(2) = \frac{3!}{(3-2)!\,2!} 0.5^{1} \times 0.5^{2}$$

$$= \frac{3 \times 2 \times 1}{1 \times (2 \times 1)} \times 0.5 \times 0.25$$

$$= \frac{6}{2} \times 0.125 \quad = \quad .375$$

Similarly, the probability of getting one toss turn up heads is

$$P(1) = \frac{3!}{(3-1)!\,1!} 0.5^{2} \times 0.5 = \frac{6}{2} \times 0.25 \times 0.5 = .375$$

and the probability of getting zero heads turn up in three tosses is

$$P(0) = \frac{3!}{(3-0)!\,0!} 0.5^{0} \times 0.5^{3} = \frac{6}{6} \times 1.0 \times 0.125 = 0.125$$

The probability of getting 2 or fewer heads in three tosses is the sum of the three probabilities, that is, 0.375 + 0.375 + 0.125 = 0.875 .

Runs Test

Random sampling is a major assumption of nearly all statistical tests of hypotheses. The Runs test is one method available for testing whether or not an obtained sample is likely to have drawn at random. It is based on the order of the values in the sample and the number of values increasing or decreasing in a sequence. For example, if a variable is composed of dichotomous values such as zeros (0) and ones (1) then a run of values such as 0,0,0,0,1,1,1,1 would not likely to have been selected at random. As another example, the values 0,1,0,1,0,1,0,1 show a definite cyclic pattern and also would not likely be found by random sampling. The test involves finding the mean of the values and examining values above and below the mean (excluding values at the mean.) The values falling above or below the mean should occur in a random fashion. A run consists of a series of values above the mean or below the mean. The expected value for the total number of runs is known and is a function of the sample size (N) and the numbers of values above (N1) and below (N2) the mean. This test may be applied to nominal through ratio variable types.

Kendall's Tau and Partial Tau

When two variables are at least ordinal, the tau correlation may be obtained as a measure of the relationship between the two variables. The values of the two variables are ranked. The method involves ordering the values using one of the variables. If the values of the other variable are in the same order, the correlation would be 1.0. If the order is exactly the opposite for this second variable, the correlation would be -1.0 just as if we had used the Pearson Product-Moment correlation method. Each pair of ranks for the second variable are compared. If the order (from low to high) is correct for a pair it is assigned a value of +1. If the pair is in reverse order, it is assigned a value of -1. These values are summed. If there are N values then we can obtain the number of pairs of scores for one variable as the number of combinations of N things taken 2 at a time which is N(N-1). The tau statistic is the ratio of the sum of 1's and -1's to the total number of pairs. Adjustments are made in the case of tied scores. For samples larger than 10, tau is approximately normally distributed.

Whenever two variables are correlated, the relationship observed may, in part, be due to their common relationship to a third variable. We may be interested in knowing what the relationship is if we partial out this third variable. The Partial Tau provides this. Since the distribution of the partial tau is not known, no test of significance is included.

The Kaplan-Meier Survival Test

Survival analysis is concerned with studying the occurrence of an event such as death or change in a subject or object at various times following the beginning of the study. Survival curves show the percentage of subjects surviving at various times as the study progresses. In many cases, it is desired to compare survival of an experimental treatment with a control treatment. This method is heavily used in medical research but is not restricted to that field. For example, one might compare the rate of college failure among students in an experimental versus a control group.

Kolmogorov-Smirnov Test

One often is interested in comparing a distribution of observed values with a theoretical distribution of values. Because many statistical tests assume a “normal” distribution, a variety of tests have been developed to determine whether or not two distributions are different beyond that expected due to random sampling variations. This test lets you compare your distribution with several theoretical distributions.

Contingency Chi-Square

The frequency chi-square statistic is used to accept or reject hypotheses concerning the degree to which observed frequencies depart from theoretical frequencies in a row by column contingency table with fixed marginal frequencies. It therefore tests the independence of the categorical variables defining the rows and columns. As an example, assume 50 males and 50 females are randomly assigned to each of three types of instructional methods to learn beginning French, (a) using a language laboratory, (b) using a computer with voice synthesizer and (c) using an advanced student tutor. Following a treatment period, a test is administered to each student with scoring results being pass or fail. The frequency of passing is then recorded for each cell in the 2 by 3 array (gender by treatment). If gender is independent of the treatment variable, the expected frequency of males that pass in each treatment would be the same as the expected frequency for females. The chi-squared statistic is obtained as

$$\chi^2 = \frac{\sum_{i=1}^{row} \sum_{j=1}^{col} (f_{ij} - F_{ij})^2}{F_{ij}}$$

where f_{ij} is the observed frequency, F_{ij} the expected frequency, and χ^2 is the chi-squared statistic with degrees of freedom (rows - 1) times (columns - 1).

The dialog for specifying a chi square analysis is shown below:

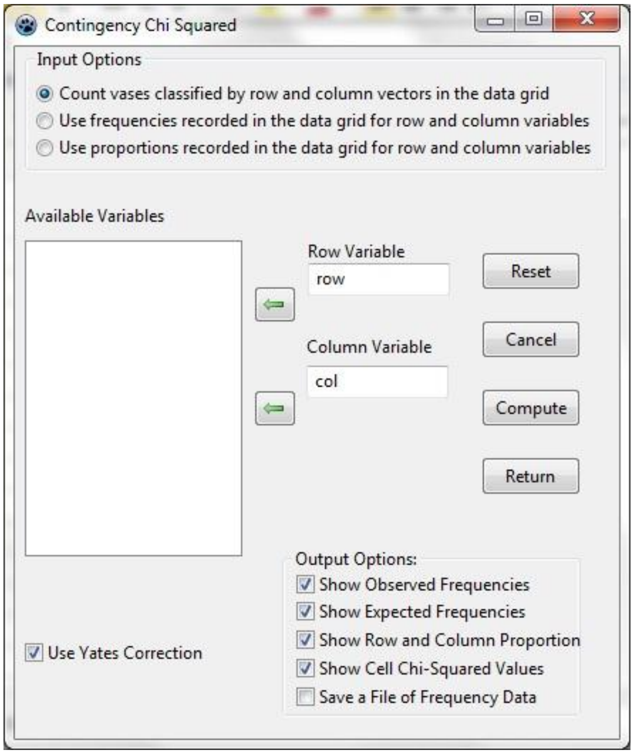


Fig. 8.1 Chi-Squared Dialog

The File ChiSqr.LAZ has been loaded for this example. When the Compute button is clicked, the following results are obtained:

```
Chi-square Analysis Results


OBSERVED FREQUENCIES


                        Rows
Variables
                 COL.1       COL.2       COL.3       COL.4       Total
     Row 1           5           5           5           5          20
     Row 2          10           4           7           3          24
     Row 3           5          10          10           2          27
     Total          20          19          22          10          71
EXPECTED FREQUENCIES with   71 cases.


Variables
                  COL.1       COL.2       COL.3       COL.4
     Row 1       5.634       5.352       6.197       2.817
     Row 2       6.761       6.423       7.437       3.380
     Row 3       7.606       7.225       8.366       3.803

ROW PROPORTIONS with   71 cases.


Variables
                  COL.1       COL.2       COL.3       COL.4       Total
     Row 1       0.250       0.250       0.250       0.250       1.000
     Row 2       0.417       0.167       0.292       0.125       1.000
     Row 3       0.185       0.370       0.370       0.074       1.000
     Total       0.282       0.268       0.310       0.141       1.000
```

```
COLUMN PROPORTIONS with   71 cases.


Variables
                  COL.1        COL.2        COL.3        COL.4        Total
     Row 1       0.250        0.263        0.227        0.500        0.282
     Row 2       0.500        0.211        0.318        0.300        0.338
     Row 3       0.250        0.526        0.455        0.200        0.380
     Total       1.000        1.000        1.000        1.000        1.000



PROPORTIONS OF TOTAL N with   71 cases.


Variables
                  COL.1        COL.2        COL.3        COL.4        Total
     Row 1       0.070        0.070        0.070        0.070        0.282
     Row 2       0.141        0.056        0.099        0.042        0.338
     Row 3       0.070        0.141        0.141        0.028        0.380
     Total       0.282        0.268        0.310        0.141        1.000

CHI-SQUARED VALUE FOR CELLS with   71 cases.


Variables
                  COL.1        COL.2        COL.3        COL.4
     Row 1       0.071        0.023        0.231        1.692
     Row 2       1.552        0.914        0.026        0.043
     Row 3       0.893        1.066        0.319        0.855
Chi-square =    7.684 with D.F. = 6. Prob. > value =    0.262

Liklihood Ratio =    7.498 with prob. > value = 0.2772

G statistic =    7.498 with prob. > value = 0.2772

phi correlation = 0.3290

Pearson Correlation r = -0.0537

Mantel-Haenszel Test of Linear Association =    0.202 with probability > value = 0.6532

The coefficient of contingency =    0.312

Cramers V =    0.233
```

Spearman Rank Correlation

When the researcher's data represent ordinal measures such as ranks with some observations being tied for the same rank, the Rank Correlation may be the appropriate statistic to calculate. While the computation for the case of untied cases is the same as that for the Pearson Product-Moment correlation, the correction for tied ranks is found only in the Spearman correlation. In addition, the interpretation of the significance of the Rank Correlation may differ from that of the Pearson Correlation where bivariate normalcy is assumed. Shown below is an example for obtaining the Spearman Rank Correlation:

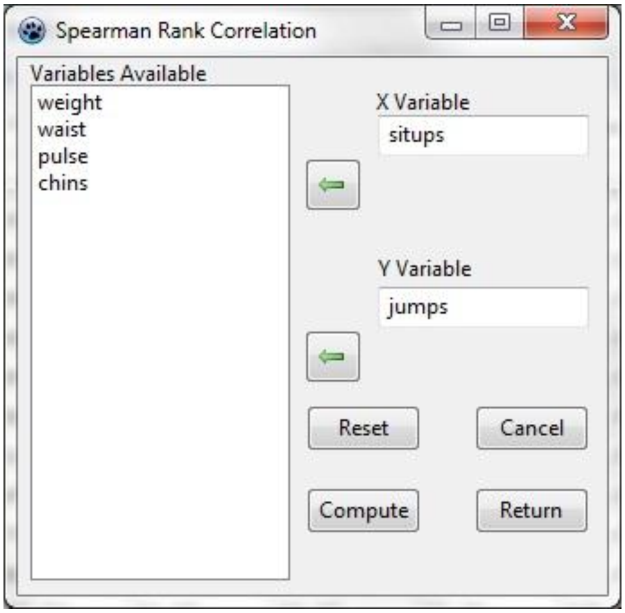


Fig. 8.2 Spearman Rank Correlation Form

```
Spearman Rank Correlation Between situps & jumps

Tied ranks correction for X =   662.00 for 0 ties
Tied ranks correction for Y =   663.50 for 0 ties

Observed scores, their ranks and differences between ranks
CASE     situps    Ranks      jumps     Ranks Rank Difference
   1     162.00     13.00     60.00     12.50      0.50
   2     110.00      8.50     60.00     12.50     -4.00
   3     101.00      5.00    101.00     16.00    -11.00
   4     105.00      7.00     37.00      3.00      4.00
   5     155.00     12.00     58.00     11.00      1.00
   6     101.00      5.00     42.00      8.00     -3.00
   7     101.00      5.00     38.00      4.50      0.50
   8     125.00     11.00     40.00      6.50      4.50
   9     200.00     14.00     40.00      6.50      7.50
  10     251.00     20.00    250.00     20.00      0.00
  11     120.00     10.00     38.00      4.50      5.50
  12     210.00     15.50    115.00     18.00     -2.50
  13     215.00     17.00    105.00     17.00      0.00
  14      50.00      1.00     50.00     10.00     -9.00
  15      70.00      3.00     31.00      2.00      1.00
  16     210.00     15.50    120.00     19.00     -3.50
  17      60.00      2.00     25.00      1.00      1.00
  18     230.00     19.00     80.00     15.00      4.00
  19     225.00     18.00     73.00     14.00      4.00
  20     110.00      8.50     43.00      9.00     -0.50
Spearman Rank Correlation =  0.695

t-test value for hypothesis r = 0 is 4.103
Probability > t = 0.0007

Pearson r for original scores :=  0.669
For the Original Scores:
Mean X  Variance X  Std.Dev. X  Mean Y  Variance Y  Std.Dev. Y
  145.55   3914.58     62.57     70.30   2629.38     51.28
```

Mann-Whitney U Test

An alternative to the Student t-test when the scale of measurement cannot be assumed to be interval or ratio and the distribution of errors is unknown is a non-parametric test known as the Mann-Whitney test. In this test, the dependent variable scores for both groups are ranked and the number of times that one groups scores exceed the rank of scores in the other group are recorded. This total number of times scores in one group exceed those of the other is named U. The sampling distribution of U is known and forms the basis for the hypothesis that the scores come from the same population.

The example below illustrates the calculation of the U test with LazStats using the mannwhitU.LAZ file:

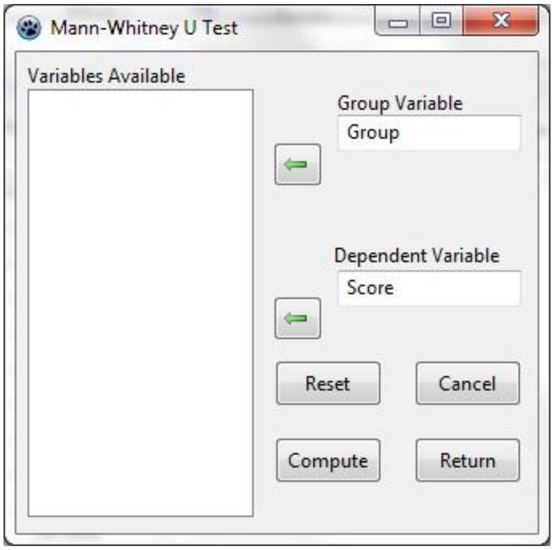


Fig. 8.3 The Mann-Whitney U Test Form

```
Mann-Whitney U Test
See pages 116-127 in S. Siegel: Nonparametric Statistics for the Behavioral Sciences

     Score      Rank       Group

      6.00       1.50          1
      6.00       1.50          2
      7.00       5.00          1
      7.00       5.00          1
      7.00       5.00          1
      7.00       5.00          1
      7.00       5.00          1
      8.00       9.50          1
      8.00       9.50          2
      8.00       9.50          2
      8.00       9.50          1
      9.00      12.00          1
     10.00      16.00          1
     10.00      16.00          2
     10.00      16.00          2
     10.00      16.00          2
     10.00      16.00          1
     10.00      16.00          1
     10.00      16.00          1
     11.00      20.50          2
     11.00      20.50          2
     12.00      24.50          2
     12.00      24.50          2
     12.00      24.50          2
     12.00      24.50          2
     12.00      24.50          1
     12.00      24.50          1
     13.00      29.50          1
     13.00      29.50          2
     13.00      29.50          2
     13.00      29.50          2
     14.00      33.00          2
     14.00      33.00          2
     14.00      33.00          2
     15.00      36.00          2
     15.00      36.00          2
     15.00      36.00          2
```

```
     16.00      38.00          2
     17.00      39.00          2

Sum of Ranks in each Group
Group   Sum    No. in Group
  1      200.00    16
  2      580.00    23

No. of tied rank groups =   9
Statistic U = 304.0000
z Statistic (corrected for ties) =   3.4262, Prob. > z = 0.0003
```

Fisher's Exact Test

Assume you have collected data on principals and superintendents concerning their agreement or disagreement to the statement "high school athletes observed drinking or using drugs should be barred from further athletic competition". You record their responses in a table as below:

	Disagree	Agree
Superintendents	2	8
Principals	4	5

You ask, are the responses of superintendents and principals significantly different? Another way to ask the question is, "what is the probability of getting the pattern of responses observed *or a more extreme pattern*?". The probability of any given pattern of responses in this 2 by 2 table may be calculated from the hypergeometric probability distribution as

$$P = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

where A, B, C, and D correspond to the frequencies in the four quadrants of the table and N corresponds to the total number of individuals sampled.

When you elect the Statistics / NonParametric / Fisher's Exact Test option from the menu, you are shown a specification form which provides for four different formats for entering data. We have elected the last format (entry of frequencies on the form itself):

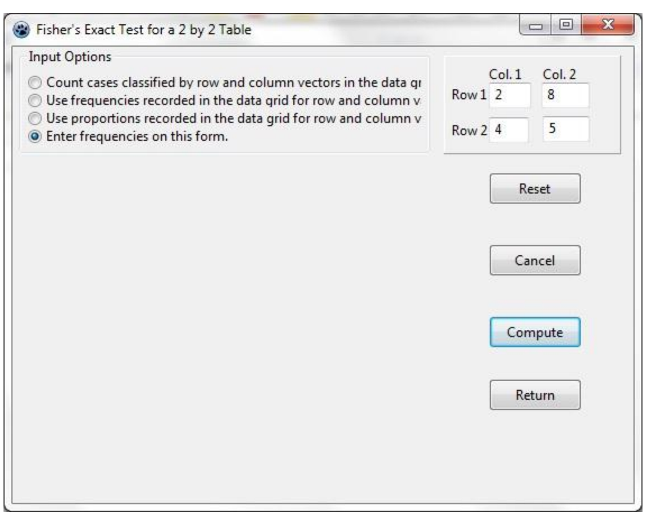


Fig. 8.4 Fisher's Exact Test Form

When we click the Compute button we obtain:

```
Fisher Exact Probability Test


Contingency Table for Fisher Exact Test
                 Column
Row             1          2
 1              2          8
 2              4          5
Probability := 0.2090

Cumulative Probability := 0.2090


Contingency Table for Fisher Exact Test
                 Column
Row             1          2
 1              1          9
 2              5          4
Probability := 0.0464

Cumulative Probability := 0.2554


Contingency Table for Fisher Exact Test
                 Column
Row             1          2
 1              0         10
 2              6          3
Probability := 0.0031

Cumulative Probability := 0.2585

Tocher ratio computed: 0.002
A random value of 0.099 selected was greater than the Tocher value.
Conclusion: Accept the null Hypothesis
```

Notice that the probability of each combination of cell values as extreme or more extreme than that observed is computed and the probabilities summed.

Kendall's Coefficient of Concordance

It is not uncommon that a group of people are asked to judge a group of persons or objects by rank ordering them from highest to lowest. It is then desirable to have some index of the degree to which the various judges agreed, that is, ranked the objects in the same order. The Coefficient of Concordance is a measure varying between 0 and 1 that indicates the degree of agreement among judges. It is defined as:

W = Variance of rank sums / maximum variance of rank sums.

The coefficient W may also be used to obtain the average rank correlation among the judges by the formula:

$$M_r = (mW - 1) / (m - 1)$$

where M_r is the average (Spearman) rank correlation, m is the number of judges and W is the Coefficient of Concordance.

The file sucsintv.LAZ is used to demonstrate the LazStats procedure for this analysis:

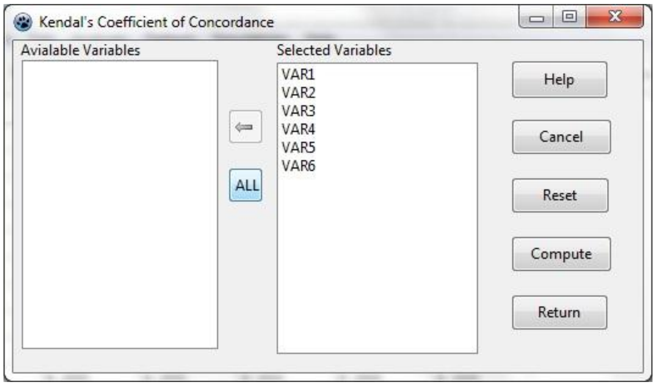


Fig. 8.5 Coefficient of Concordance Form

```
Kendall Coefficient of Concordance Analysis

Ranks Assigned to Judge Ratings of Objects

Judge    1              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   5.5000  4.0000  3.0000  2.0000  1.0000  5.5000

Judge    2              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  4.5000  3.0000  2.0000  1.0000  4.5000

Judge    3              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   5.5000  3.5000  5.5000  2.0000  1.0000  3.5000

Judge    4              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  5.0000  3.5000  2.0000  1.0000  3.5000

Judge    5              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  4.0000  4.0000  2.0000  1.0000  4.0000

Judge    6              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  5.0000  4.0000  2.5000  1.0000  2.5000

Judge    7              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  5.0000  4.0000  3.0000  2.0000  1.0000

Judge    8              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  4.5000  3.0000  2.0000  1.0000  4.5000

Judge    9              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  4.5000  3.0000  2.0000  1.0000  4.5000

Judge   10              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  5.0000  3.5000  2.0000  1.0000  3.5000

Judge   11              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  4.5000  2.5000  4.5000  1.0000  2.5000

Judge   12              Objects
    VAR1     VAR2     VAR3     VAR4     VAR5      VAR6
   6.0000  5.0000  4.0000  2.5000  1.0000  2.5000


Sum of Ranks for Each Object Judged
```

```
             Objects
    VAR1    VAR2    VAR3    VAR4    VAR5     VAR6
  71.0000 54.5000 43.0000 28.5000 13.0000 42.0000

Coefficient of concordance :=      0.834
Average Spearman Rank Correlation :=      0.819
Chi-Square Statistic :=   50.037
Probability of a larger Chi-Square := 0.0000
Warning - Above Chi-Square is very approximate with 7 or fewer variables!
```

Kruskal-Wallis One-Way ANOVA

One-Way, Fixed-Effects Analysis of Variance assumes that error (residual) scores are normally distributed, that subjects are randomly selected from the population and assigned to treatments, and that the error scores are equally distributed in the populations representing the treatments. The scale of measurement for the dependent variable is assumed to be interval or ratio. But what can you do if, in fact, your measure is only ordinal (for example like most classroom tests), and you cannot assume normally distributed, homoscedastic error distributions?

Why, of course, you convert the scores to ranks and ask if the sum of rank scores in each treatment group are the same within sampling error! The Kruskal-Wallis One-Way Analysis of variance converts the dependent score for each subject in the study to a rank from 1 to N. It then examines the ranks attained by subjects in each of the treatment groups. Then a test statistic which is distributed as Chi-Squared with degrees of freedom equal to the number of treatment groups minus one is obtained from:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{K} R_j^2 / n_j - 3(N+1)$$

where N is the total number of subjects in the experiment, n_j is the number of subjects in the jth treatment, K is the number of treatments and R_j is the sum of ranks in the jth treatment.

The Statistics / NonParametric / Kruskal-Wallis One-Way ANOVA option on your menu will permit analysis of data in a data file. At least two variables should be defined for each case - a variable recording the treatment group for the case and a variable containing the dependent variable score.

The file labeled kwanova.LAZ is used to demonstrate this LazStats procedure.

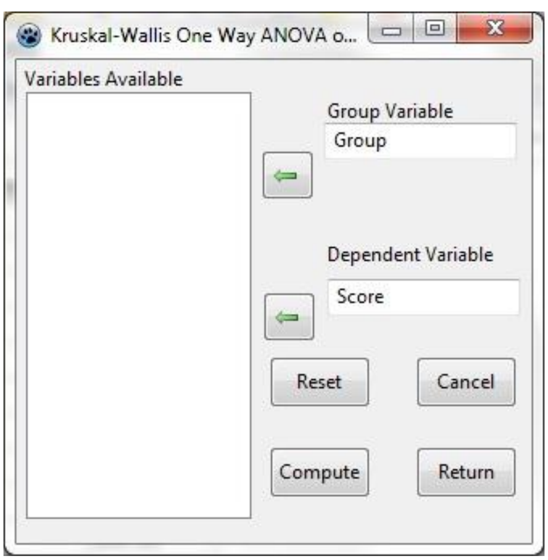


Fig. 8.6 The Kruskal-Wallis One-Way ANOVA Form

```
Kruskal - Wallis One-Way Analysis of Variance
```

```
See pages 184-194 in S. Siegel: Nonparametric Statistics for the Behavioral Sciences

     Score     Rank      Group

     61.00      1.00         1
     82.00      2.00         2
     83.00      3.00         1
     96.00      4.00         1
    101.00      5.00         1
    109.00      6.00         2
    115.00      7.00         3
    124.00      8.00         2
    128.00      9.00         1
    132.00     10.00         2
    135.00     11.00         2
    147.00     12.00         3
    149.00     13.00         3
    166.00     14.00         3

Sum of Ranks in each Group
Group   Sum    No. in Group
  1      22.00     5
  2      37.00     5
  3      46.00     4

No. of tied rank groups =   0
Statistic H uncorrected for ties =   6.4057
Correction for Ties = 1.0000
Statistic H corrected for ties =   6.4057
Corrected H is approx. chi-square with   2 D.F. and probability = 0.0406
```

Wilcoxon Matched-Pairs Signed Ranks Test

This test provides an alternative to the student t-test for matched score data where the assumptions for the parametric t-test cannot be met. In using this test, the difference is obtained between each of N pairs of scores observed on matched objects, for example, the difference between pretest and post-test scores for a group of students. The difference scores obtained are then ranked. The ranks of negative score differences are summed and the ranks of positive score differences are summed. The test statistic T is the smaller of these two sums. Difference scores of 0 are eliminated since a rank cannot be assigned. If the null hypothesis of no difference between the groups of scores is true, the sum of positive ranks should not differ from the sum of negative ranks beyond that expected by chance. Given N ranks, there is a finite number of ways of obtaining a given sum T. There are a total of 2 raised to the N ways of assigning positive and negative differences to N ranks. In a sample of 5 pairs, for example, there are 2 to the fifth power = 32 ways. Each rank sign would occur with probability of 1/32. The probability of getting a particular total T is

$$P_T = \frac{\text{Ways of getting T}}{2^N}$$

The cumulative probabilities for T, T-1,....,0 are obtained for the observed T value and reported. For large samples, a normally distributed z score is approximated and used.

The file labeled wilcoxon.LAZ is used as our example:

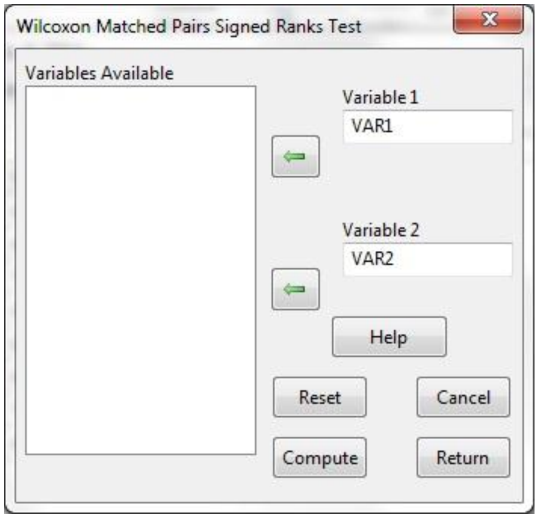


Fig. 8.7 The Wilcoxon Matched Pairs Signed Ranks Test

```
The Wilcoxon Matched-Pairs Signed-Ranks Test
See pages 75-83 in S. Seigel's Nonparametric Statistics for the Social Sciences

Ordered Cases with cases having 0 differences eliminated:
Number of cases with absolute differences greater than 0 = 8
CASE      VAR1         VAR2    Difference   Signed Rank
  3       73.00       74.00     -1.00     -1.00
  8       65.00       62.00      3.00      2.00
  7       76.00       80.00     -4.00     -3.00
  4       43.00       37.00      6.00      4.00
  5       58.00       51.00      7.00      5.00
  6       56.00       43.00     13.00      6.00
  1       82.00       63.00     19.00      7.00
  2       69.00       42.00     27.00      8.00

Smaller sum of ranks (T) =     4.00
Approximately normal z for test statistic T =  1.960
Probability (1-tailed) of greater z = 0.0250
NOTE: For N < 25 use tabled values for Wilcoxon Test
```

Cochran Q Test

The Cochran Q test is used to test whether or not two or more matched sets of frequencies or proportions, measured on a nominal or ordinal scale, differ significantly among themselves. Typically, observations are dichotomous, that is, scored as 0 or 1 depending on whether or not the subject falls into one or the other criterion group. An example of research for which the Q test may be applied might be the agreement or disagreement to the question "Should abortions be legal?". The research design might call for a sample of n subjects answering the question prior to a debate and following a debate on the topic and subsequently six months later. The Q test applied to these data would test whether or not the proportion agreeing was the same under these three time periods. The Q statistic is obtained as

$$Q = \frac{(K - 1) \sum_{j=1}^{K} G_j^2 - \left(\sum_{j=1}^{K} G_j \right)^2}{K \sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}$$

where K is the number of treatments (groups of scores, G_j is the sum with the jth treatment group, and L_i is the sum within case i (across groups). The Q statistic is distributed approximately as Chi-squared with degrees of freedom K-1. If Q exceeds the Chi-Squared value corresponding to the cumulative probability value, the hypothesis of equal proportions for the K groups is rejected. ItemData.TAB is the file used to demonstrate this procedure.

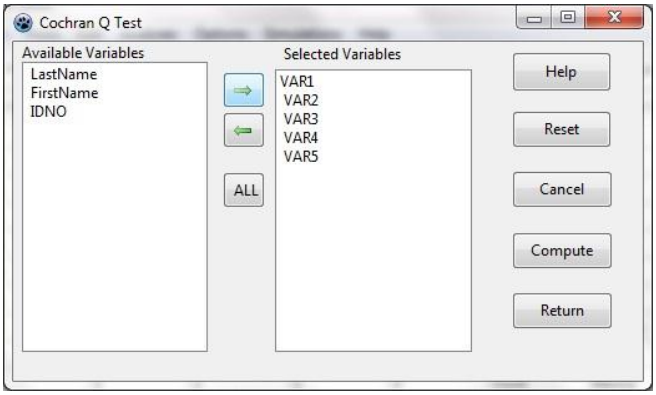


Fig. 8.8 The Q Test Form

```
Cochran Q Test for Related Samples
See pages 161-166 in S. Siegel: Nonparametric Statistics for the Behavioral Sciences
McGraw-Hill Book Company, New York, 1956

Cochran Q Statistic =  3.000
which is distributed as chi-square with 1 D.F. and probability = 0.0833
```

Sign Test

Did you hear about the nonparametrician who couln't get his driving license? He couldn't pass the sign test.

Imagine a counseling psychologist who sees, over a period of months, a number of clients with personal problems. Suppose the psychologist routinely contacts each client for a six month followup to see how they are doing. The counselor could make an estimate of client "adjustment" before treatment and at the followup time (or better still, have another person independently estimate adjustment at these two time periods). We may assume some underlying continuous "adjustment" variable even though we have no idea about the population distribution of the variable. We are intrested in knowing, of course, whether or not people are better adjusted six months after therapy than before. Note that we are only comparing the "before" and "after" state of the individuals with each other, not with other subjects. If we assign a + to the situation of improved adjustment and a - to the situation of same or poorer adjustment, we have the data required for a Sign Test. If treatment has had no effect, we would expect approximately one half the subjects would receive plus signs and the others negative signs. The sampling distribution of the proportion of plus signs is given by the binomial probability distribution with parameter of .5 and the number of events equal to n, the number of pairs of observations. We will use a file labeled signtest.tab for an example. It contains an “adjustment” score for married couples (M and F):

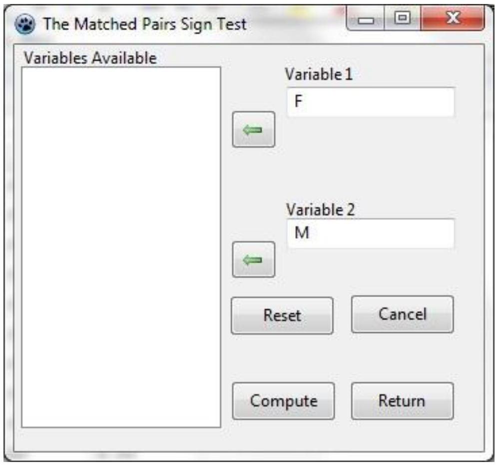


Fig. 8.9 The Sign Test Form

```
Results for the Sign Test

Frequency of  11 out of 17 observed + sign differences.
Frequency of   3 out of 17 observed - sign differences.
Frequency of   3 out of 17 observed no differences.
The theoretical proportion expected for +'s or -'s is 0.5
The test is for the probability of the +'s or -'s (which ever is fewer)
as small or smaller than that observed given the expected proportion.

Binary Probability of   0 = 0.0001
Binary Probability of   1 = 0.0009
Binary Probability of   2 = 0.0056
Binary Probability of   3 = 0.0222
Binomial Probability of     3 or smaller out of    14 = 0.0287
```

Friedman Two Way ANOVA

Imagine an experiment using, say, ten groups of subjects with four subjects in each group that have been matched on some relevant variables (or even using the same subjects). The matched subjects in each group are exposed to four different treatments such as teaching methods, dosages of medicine, proportion of positive responses to statements or questions, etc. Assume that some criterion measure on at least a nominal scale is available to measure the effect of each treatment. Now rank the subjects in each group on the basis of their scores on the criterion. We may now ask whether the ranks in each treatment come from the same population. Had we been able to assume an interval or ratio measure and normally distributed errors, we might have used a repeated measures analysis of variance. Failing to meet the parametric test assumptions, we instead examine the sum of ranks obtained under each of the treatment conditions and ask whether they differ significantly. The test statistic is distributed as Chi-squared with degrees of freedom equal to the number of treatments minus one. It is obtained as where N is the number of groups, K the number of treatments (or number of subjects in each group), and R_j is the sum of ranks in each treatment.

Friedman_ties.LAZ will be used to demonstrate this procedure. Shown below is the dialog form to specify the analysis and the results of the analysis.

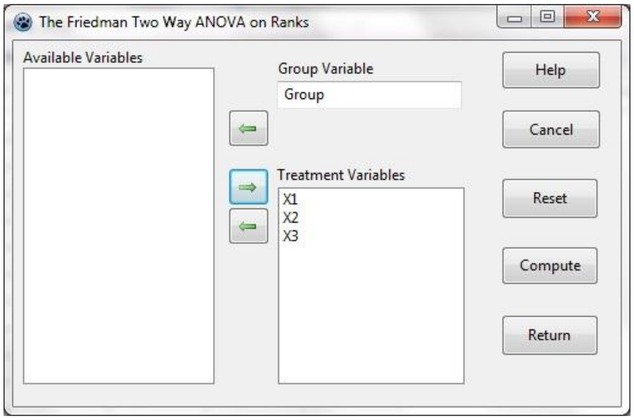


Fig. 8.10 The Friedman Analysis Specification Form

FRIEDMAN TWO-WAY ANOVA ON RANKS
See pages 166-173 in S. Siegel's Nonparametric Statistics
for the Behavioral Sciences, McGraw-Hill Book Co., New York, 1956

Treatment means - values to be ranked. with 18 cases.

Variables

	X1	X2	X3
Group 1	10.333	10.333	7.667
Group 2	11.000	10.667	7.000
Group 3	12.333	10.333	6.333
Group 4	5.333	5.667	3.667
Group 5	4.000	3.000	2.667
Group 6	3.000	2.000	2.333

Number in each group's treatment.

GROUP

Variables

	X1	X2	X3
Group 1	3	3	3
Group 2	3	3	3
Group 3	3	3	3
Group 4	3	3	3
Group 5	3	3	3
Group 6	3	3	3

Score Rankings Within Groups with 18 cases.

Variables

	X1	X2	X3
Group 1	2.500	2.500	1.000
Group 2	3.000	2.000	1.000
Group 3	3.000	2.000	1.000
Group 4	2.000	3.000	1.000
Group 5	3.000	2.000	1.000
Group 6	3.000	1.000	2.000

TOTAL RANKS with 18 valid cases.

Variables X1 X2 X3

```
                16.500       12.500        7.000


Chi-square with 2 D.F. :=    7.583 with probability := 0.0226
Chi-square too approximate-use exact table (TABLE N)
page 280-281 in Siegel
```

Probability of a Binomial Event

The BINOMIAL program is a short program to calculate the probability of obtaining k or fewer occurrences of a dichotomous variable out of a total of n observations when the probability of an occurrence is known. For example, assume a test consists of 5 multiple choice items with each item scored correct or incorrect. Also assume that there are five equally plausible choices for a student with no knowledge concerning any item. In this case, the probability of a student guessing the correct answer to a single item is 1/5 or .20 . We may use the binomial program to obtain the probabilities that a student guessing on each item of the test gets a score of 0, 1, 2, 3, 4, or 5 items correct by chance alone.

The formula for the probability of a dichotomous event k where the probability of a single event is p (and the probability of a non-event is q = 1 - p is given as:

$$P(k) = \frac{N!}{(N-k)!\,k!}\, p^{(N-k)} \quad q^{k}$$

For example, if a “fair” coin is tossed three times with the probabilities of heads is p = .5 (and q = .5) then the probabilty of observing 2 heads is

$$P(2) = \frac{3!}{(3-2)!\,2!} 0.5^{1} \times 0.5^{2}$$

$$= \frac{3 \times 2 \times 1}{1 \times (2 \times 1)} \times 0.5 \times 0.25$$

$$= \frac{6}{2} \times 0.125 \quad = \quad .375$$

Similarly, the probability of getting one toss turn up heads is

$$P(1) = \frac{3!}{(3-1)!\,1!} 0.5^{2} \times 0.5 = \frac{6}{2} \times 0.25 \times 0.5 = .375$$

and the probability of getting zero heads turn up in three tosses is

$$P(0) = \frac{3!}{(3-0)!\,0!} 0.5^{0} \times 0.5^{3} = \frac{6}{6} \times 1.0 \times 0.125 = 0.125$$

The probability of getting 2 or fewer heads in three tosses is the sum of the three probabilities, that is, 0.375 + 0.375 + 0.125 = 0.875 .

Shown below is the form used to obtain binomial probabilities and an example run of the procedure:

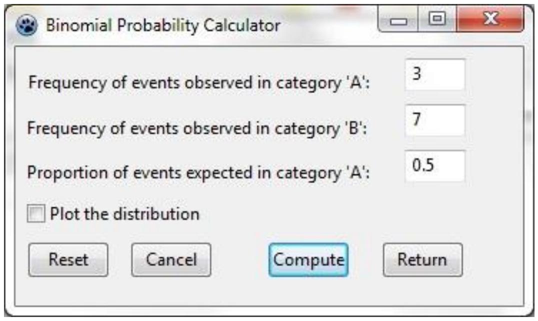


Fig. 8.11 Binomial Probability Form

```
Binomial Probability Test

Frequency of 3 out of 10 observed
The theoretical proportion expected in category A is 0.500
The test is for the probability of a value in category A as small or smaller
than that observed given the expected proportion.
Probability of 0 = 0.0010
Probability of 1 = 0.0098
Probability of 2 = 0.0439
Probability of 3 = 0.1172
Binomial Probability of 3 or less out of 10 = 0.1719
```

Runs Test

Random sampling is a major assumption of nearly all statistical tests of hypotheses. The Runs test is one method available for testing whether or not an obtained sample is likely to have been drawn at random. It is based on the order of the values in the sample and the number of values increasing or decreasing in a sequence. For example, if a variable is composed of dichotomous values such as zeros (0) and ones (1) then a run of values such as 0,0,0,0,1,1,1,1 would not likely to have been selected at random. As another example, the values 0,1,0,1,0,1,0,1 show a definite cyclic pattern and also would not likely be found by random sampling. The test involves finding the mean of the values and examining values above and below the mean (excluding values at the mean.) The values falling above or below the mean should occur in a random fashion. A run consists of a series of values above the mean or below the mean. The expected value for the total number of runs is known and is a function of the sample size (N) and the numbers of values above (N1) and below (N2) the mean. This test may be applied to nominal through ratio variable types.

The file labeled runstest.LAZ will be used to demonstrate the use of this procedure.

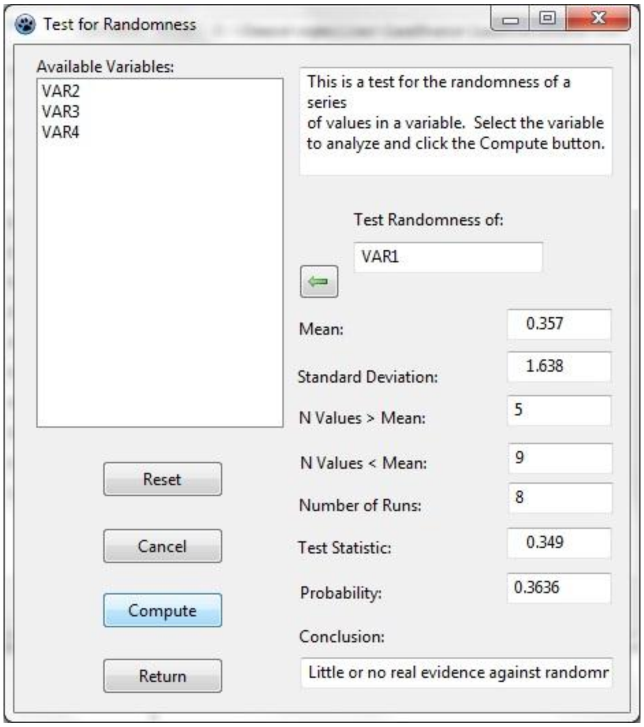


Fig. 8.12 Test for Randomness Using the Runs Test

Kendall's Tau and Partial Tau

When two variables are at least ordinal, the tau correlation may be obtained as a measure of the relationship between the two variables. The values of the two variables are ranked. The method involves ordering the values using one of the variables. If the values of the other variable are in the same order, the correlation would be 1.0. If the order is exactly the opposite for this second variable, the correlation would be -1.0 just as if we had used the Pearson Product-Moment correlation method. Each pair of ranks for the second variable are compared. If the order (from low to high) is correct for a pair it is assigned a value of +1. If the pair is in reverse order, it is assigned a value of -1. These values are summed. If there are N values then we can obtain the number of pairs of scores for one variable as the number of combinations of N things taken 2 at a time which is N(N-1). The tau statistic is the ratio of the sum of 1's and -1's to the total number of pairs. Adjustments are made in the case of tied scores. For samples larger than 10, tau is approximately normally distributed.

Whenever two variables are correlated, the relationship observed may, in part, be due to their common relationship to a third variable. We may be interested in knowing what the relationship is if we partial out this third variable. The Partial Tau provides this. Since the distribution of the partial tau is not known, no test of significance is included.

The file labeled TAUDATA.LAZ has been used to illustrate this procedure in the Fig. below:

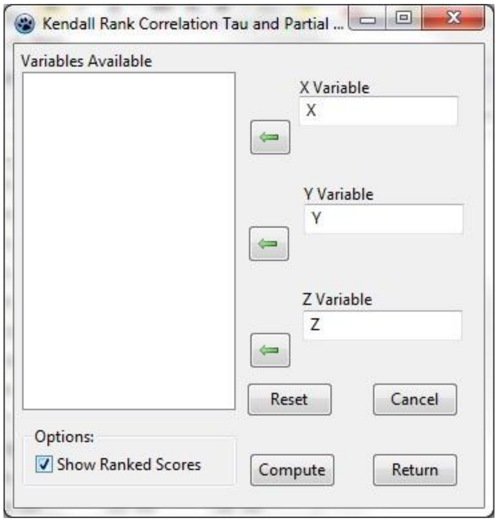


Fig. 8.13 Kendall's Tau and Partial Tau Form

```
Ranks with   12 cases.


Variables
                      X            Y             Z
         1       3.000        2.000         1.500
         2       4.000        6.000         1.500
         3       2.000        5.000         3.500
         4       1.000        1.000         3.500
         5       8.000       10.000         5.000
         6      11.000        9.000         6.000
         7      10.000        8.000         7.000
         8       6.000        3.000         8.000
         9       7.000        4.000         9.000
        10      12.000       12.000        10.500
        11       5.000        7.000        10.500
        12       9.000       11.000        12.000

Kendall Tau for File: C:\lazarus\Projects\LazStats\LazStatsData\TAUDATA.LAZ

Kendall Tau for variables X and Y
Tau =   0.6667  z =    3.017 probability > |z| = 0.001

Kendall Tau for variables X and Z
Tau =   0.3877  z =    1.755 probability > |z| = 0.040

Kendall Tau for variables Y and Z
Tau =   0.3567  z =    1.614 probability > |z| = 0.053

Partial Tau =   0.6136

NOTE: Probabilities are for large N (>10)
```

The Kaplan-Meier Survival Test

Survival analysis is concerned with studying the occurrence of an event such as death or change in a subject or object at various times following the beginning of the study. Survival curves show the percentage of subjects surviving at various times as the study progresses. In many cases, it is desired to compare survival of an experimental treatment with a control treatment. This method is heavily used in medical research but is not restricted to that field. For example, one might compare the rate of college failure among students in an experimental versus a control group.

To obtain a survival curve you need only two items of information in your data file for each subject: the survival time and a code for whether or not an event occurred or the subject has been lost from the study moved, disappeared, etc. (censored.) If an event such as death occurred, it is coded as a 1. If censored it is coded as a 2.

```
CASES FOR FILE C:\LazStats\KaplanMeier.LAZ

           0        Time  Event_Censored
           1           1          2
           2           3          2
           3           5          2
           4           6          1
           5           6          1
           6           6          1
           7           6          1
           8           6          1
           9           6          1
          10           8          1
          11           8          1
          12           9          2
          13          10          1
          14          10          1
          15          10          2
          16          12          1
          17          12          1
          18          12          1
          19          12          1
          20          12          1
          21          12          1
          22          12          2
          23          12          2
          24          13          2
          25          15          2
          26          15          2
          27          16          2
          28          16          2
          29          18          2
          30          18          2
          31          20          1
          32          20          2
          33          22          2
          34          24          1
          35          24          1
          36          24          2
          37          27          2
          38          28          2
          39          28          2
          40          28          2
          41          30          1
          42          30          2
          43          32          1
          44          33          2
          45          34          2
          46          36          2
          47          36          2
          48          42          1
          49          44          2
```

We are really recording data for the "Time" variable that is sequential through the data file. We are concerned with the percent of survivors at any given time period as we progress through the observation times of the study. We record the "drop-outs" or censored subjects at each time period also. A unit cannot be censored and be one of the deaths - these are mutually exclusive.

Next we show a data file that contains both experimental and control subjects:

```
CASES FOR FILE C:\LazStats\KaplanMeier.LAZ

           0        Time       Group  Event_Censored
           1           1           1           2
           2           3           2           2
           3           5           1           2
```

4	6	1	1
5	6	1	1
6	6	2	1
7	6	2	1
8	6	2	1
9	6	2	1
10	8	2	1
11	8	2	1
12	9	1	2
13	10	1	1
14	10	1	1
15	10	1	2
16	12	1	1
17	12	1	1
18	12	1	1
19	12	1	1
20	12	2	1
21	12	2	1
22	12	1	2
23	12	2	2
24	13	1	2
25	15	1	2
26	15	2	2
27	16	1	2
28	16	2	2
29	18	2	2
30	18	2	2
31	20	2	1
32	20	1	2
33	22	2	2
34	24	1	1
35	24	2	1
36	24	1	2
37	27	1	2
38	28	2	2
39	28	2	2
40	28	2	2
41	30	2	1
42	30	2	2
43	32	1	1
44	33	2	2
45	34	1	2
46	36	1	2
47	36	1	2
48	42	2	1
49	44	1	2

In this data we code the groups as 1 or 2. Censored cases are always coded 2 and Events are coded 1. This data is, in fact, the same data as shown in the previous data file. Note that in time period 6 there were 6 deaths (cases 4-9.) Again, notice that the time periods are in ascending order.

Shown below is the specification dialog for this second data file. This is followed by the output obtained when you click the compute button.

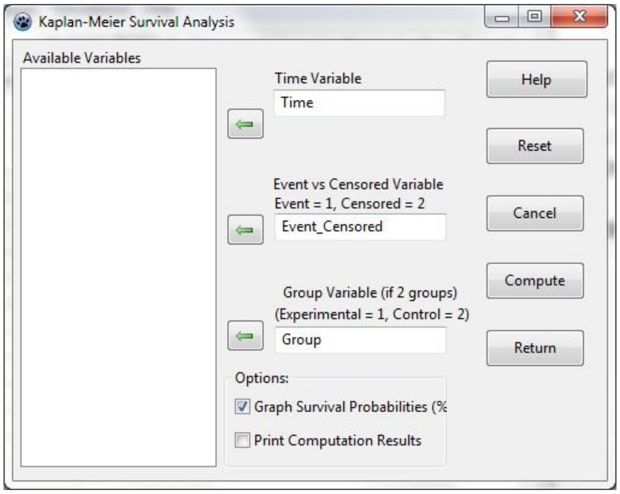


Fig. 8.14 The Kaplan-Meier Dialog

```
Kaplan-Meier Survival Test

Comparison of Two Groups Methd
```

TIME	GROUP	CENSORED	TOTAL AT RISK	EVENTS	AT RISK IN GROUP 1	EXPECTED NO. EVENTS IN 1	AT RISK IN GROUP 2	EXPECTED NO. EVENTS IN 2
0	0	0	49	0	25	0	24	0
1	1	1	49	0	25	0	24	0
3	2	1	48	0	24	0	24	0
5	1	1	47	0	24	0	23	0
6	1	0	46	6	23	3	23	3
6	1	0	40	0	21	0	19	0
6	2	0	40	0	21	0	19	0
6	2	0	40	0	21	0	19	0
6	2	0	40	0	21	0	19	0
6	2	0	40	0	21	0	19	0
8	2	0	40	2	21	2	19	1
8	2	0	38	0	21	0	17	0
9	1	1	38	0	21	0	17	0
10	1	0	37	2	20	2	17	1
10	1	0	35	0	18	0	17	0
10	1	1	35	0	18	0	17	0
12	1	0	34	6	17	3	17	3
12	1	0	28	0	13	0	15	0
12	1	0	28	0	13	0	15	0
12	1	0	28	0	13	0	15	0
12	2	0	28	0	13	0	15	0
12	2	0	28	0	13	0	15	0
12	1	1	28	0	13	0	15	0
12	2	1	27	0	12	0	15	0
13	1	1	26	0	12	0	14	0
15	1	1	25	0	11	0	14	0
15	2	1	24	0	10	0	14	0
16	1	1	23	0	10	0	13	0
16	2	1	22	0	9	0	13	0
18	2	1	21	0	9	0	12	0
18	2	1	20	0	9	0	11	0
20	2	0	19	1	9	1	10	1
20	1	1	18	0	9	0	9	0
22	2	1	17	0	8	0	9	0
24	1	0	16	2	8	1	8	1
24	2	0	14	0	7	0	7	0
24	1	1	14	0	7	0	7	0
27	1	1	13	0	6	0	7	0
28	2	1	12	0	5	0	7	0
28	2	1	11	0	5	0	6	0
28	2	1	10	0	5	0	5	0
30	2	0	9	1	5	1	4	1
30	2	1	8	0	5	0	3	0
32	1	0	7	1	5	1	2	1

```
  33    2        1        6        0        4          0          2            0
  34    1        1        5        0        4          0          1            0
  36    1        1        4        0        3          0          1            0
  36    1        1        3        0        2          0          1            0
  42    2        0        2        1        1          1          1            1
  44    1        1        0        0        1          0          0            0

TIME  DEATHS  GROUP  AT RISK  PROPORTION  CUMULATIVE
                              SURVIVING   PROP.SURVIVING
   1      0      1       25      0.0000       1.0000
   3      0      2       24      0.0000       1.0000
   5      0      1       24      0.0000       1.0000
   6      6      1       23      0.9130       0.9130
   6      0      1       21      0.0000       0.9130
   6      0      2       19      0.0000       0.8261
   6      0      2       19      0.0000       0.8261
   6      0      2       19      0.0000       0.8261
   6      0      2       19      0.0000       0.8261
   8      2      2       19      0.8947       0.7391
   8      0      2       17      0.0000       0.7391
   9      0      1       21      0.0000       0.9130
  10      2      1       20      0.9000       0.8217
  10      0      1       18      0.0000       0.8217
  10      0      1       18      0.0000       0.8217
  12      6      1       17      0.7647       0.6284
  12      0      1       13      0.0000       0.6284
  12      0      1       13      0.0000       0.6284
  12      0      1       13      0.0000       0.6284
  12      0      2       15      0.0000       0.6522
  12      0      2       15      0.0000       0.6522
  12      0      1       13      0.0000       0.6284
  12      0      2       15      0.0000       0.6522
  13      0      1       12      0.0000       0.6284
  15      0      1       11      0.0000       0.6284
  15      0      2       14      0.0000       0.6522
  16      0      1       10      0.0000       0.6284
  16      0      2       13      0.0000       0.6522
  18      0      2       12      0.0000       0.6522
  18      0      2       11      0.0000       0.6522
  20      1      2       10      0.9000       0.5870
  20      0      1        9      0.0000       0.6284
  22      0      2        9      0.0000       0.5870
  24      2      1        8      0.8750       0.5498
  24      0      2        7      0.0000       0.5136
  24      0      1        7      0.0000       0.5498
  27      0      1        6      0.0000       0.5498
  28      0      2        7      0.0000       0.5136
  28      0      2        6      0.0000       0.5136
  28      0      2        5      0.0000       0.5136
  30      1      2        4      0.7500       0.3852
  30      0      2        3      0.0000       0.3852
  32      1      1        5      0.8000       0.4399
  33      0      2        2      0.0000       0.3852
  34      0      1        4      0.0000       0.4399
  36      0      1        3      0.0000       0.4399
  36      0      1        2      0.0000       0.4399
  42      1      2        1      0.0000       0.0000
  44      0      1        1      0.0000       0.4399

Total Expected Events for Experimental Group :=   15.000
Observed Events for Experimental Group :=   10.000
Total Expected Events for Control Group :=    7.000
Observed Events for Control Group :=   12.000
Chisquare =     5.238 with probability = 0.978
Risk :=    0.389, Log Risk :=   -0.944, Std.Err. Log Risk :=    0.458
95 Percent Confidence interval for Log Risk = (-1.842,-0.047)
95 Percent Confidence interval for Risk := (0.159,0.954)
EXPERIMENTAL GROUP CUMULATIVE PROBABILITY
CASE TIME DEATHS CENSORED CUM.PROB.
  1      1     0       1      1.000
  3      5     0       1      1.000
  4      6     6       0      0.913
  5      6     0       0      0.913
```

12	9	0	1	0.913
13	10	2	0	0.822
14	10	0	0	0.822
15	10	0	1	0.822
16	12	6	0	0.628
17	12	0	0	0.628
18	12	0	0	0.628
19	12	0	0	0.628
22	12	0	1	0.628
24	13	0	1	0.628
25	15	0	1	0.628
27	16	0	1	0.628
32	20	0	1	0.628
34	24	2	0	0.550
36	24	0	1	0.550
37	27	0	1	0.550
43	32	1	0	0.440
45	34	0	1	0.440
46	36	0	1	0.440
47	36	0	1	0.440
49	44	0	1	0.440

CONTROL GROUP CUMULATIVE PROBABILITY

CASE	TIME	DEATHS	CENSORED	CUM.PROB.
2	3	0	1	1.000
6	6	0	0	0.826
7	6	0	0	0.826
8	6	0	0	0.826
9	6	0	0	0.826
10	8	2	0	0.739
11	8	0	0	0.739
20	12	0	0	0.652
21	12	0	0	0.652
23	12	0	1	0.652
26	15	0	1	0.652
28	16	0	1	0.652
29	18	0	1	0.652
30	18	0	1	0.652
31	20	1	0	0.587
33	22	0	1	0.587
35	24	0	0	0.514
38	28	0	1	0.514
39	28	0	1	0.514
40	28	0	1	0.514
41	30	1	0	0.385
42	30	0	1	0.385
44	33	0	1	0.385
48	42	1	0	0.000

The chi-square coefficient as well as the graph indicates no difference was found between the experimental and control group beyond what is reasonably expected through random selection from the same population.

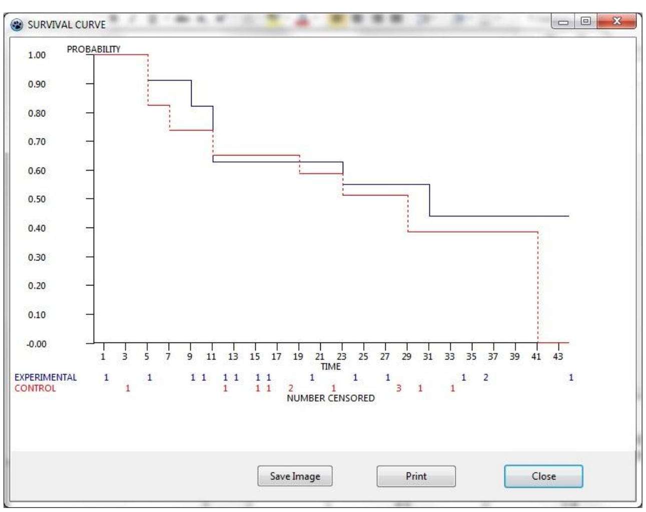


Fig. 8.15 Kaplan-Meier Survival Plot

Sen's Slope Estimate (Series Data)

The BoltSize.LAZ file is used to illustrate this procedure. The purpose is to estimate the slope from one time period to another time period for a series of data over equal intervals of time. The optional plot provides a graphical representation of the slopes obtained. One can often visually spot non-random patterns in the data and cyclic trends.

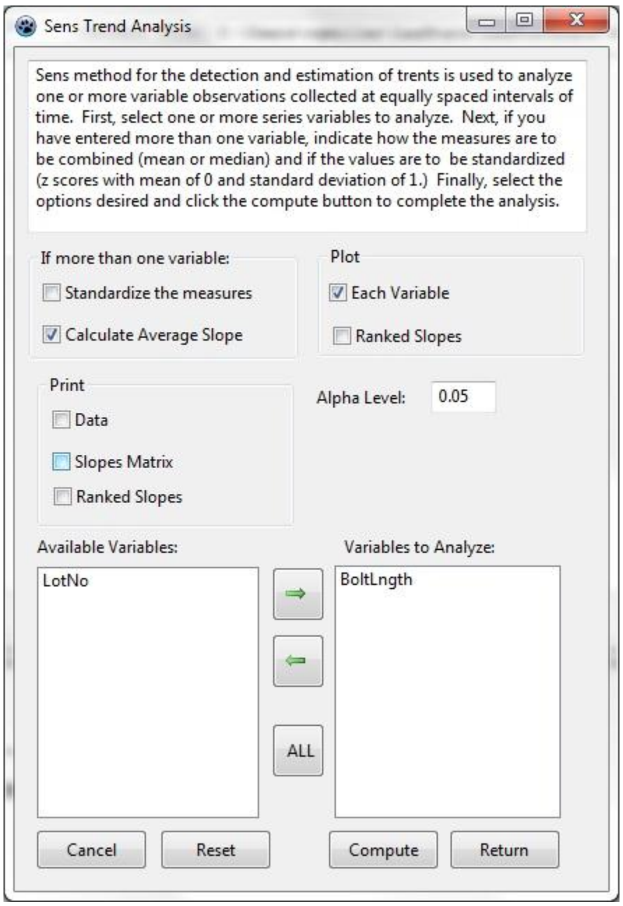


Fig. 8.16 Sen's Slope Estimates for Series Data

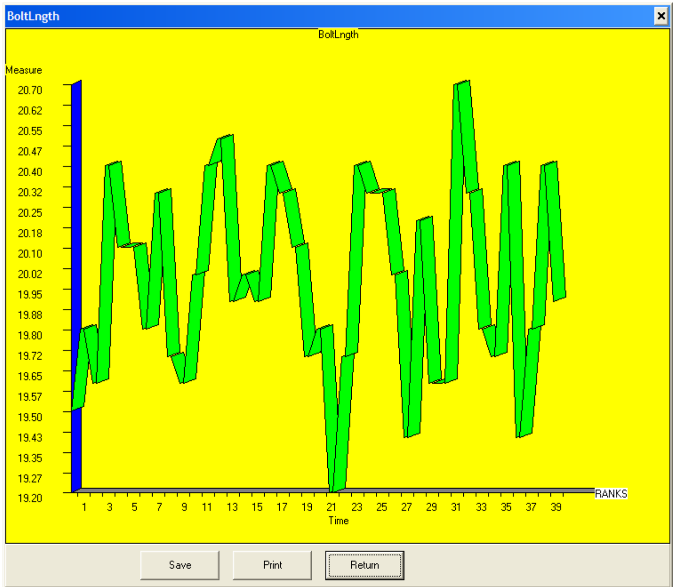


Fig. 8.17 Plot of Slopes From Sen's Slope Estimates

```
Sens Detection and Estimation of Trends
Number of data points = 40, Confidence Interval = 0.95

Results for BoltLngth
Median Slope for 780 values =    0.000
Mann-Kendall Variance statistic = 1973.000 (2 ties)
Ranks of the lower and upper confidence =  353.469,  427.531
Corresponding lower and upper slopes =    0.000,    0.000
```

Kolmogorov-Smirnov Test

One often is interested in comparing a distribution of observed values with a theoretical distribution of values. Because many statistical tests assume a “normal” distribution, a variety of tests have been developed to determine whether or not two distributions are different beyond that expected due to random sampling variations. This test lets you compare your distribution with several theoretical distributions. We have loaded the file labeled “Cansas.LAZ” as an example. Below is the dialog form used to specify our analysis. We will compare the distribution of the waist variable to a normal distribution.

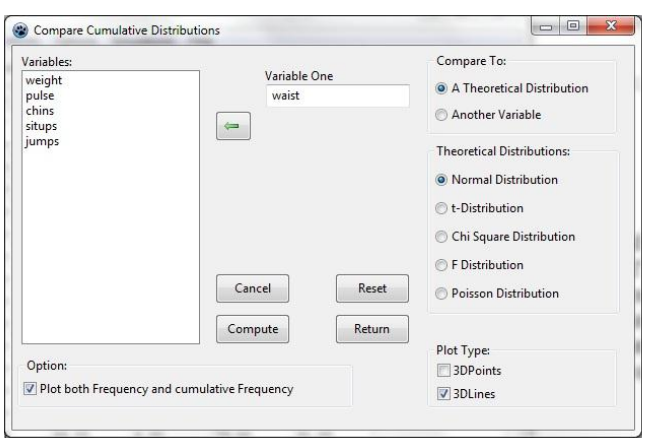


Fig. 8.18 Kolmogorov-Smirnov Test of Similar Distributions

The results for our sample are shown below:

```
Distribution comparison by Bill Miller

                                           Normal       Normal       Normal
```

```
  X1 Value     Frequency    Cum. Freq.      X2 Value     Frequency    Cum. Freq.
    31.000             1         1.000        -3.000             0         0.000
    32.000             1         2.000        -2.625             0         0.000
    33.000             4         6.000        -2.250             0         0.000
    34.000             3         9.000        -1.875             1         1.000
    35.000             2        11.000        -1.500             1         2.000
    36.000             3        14.000        -1.125             2         4.000
    37.000             3        17.000        -0.750             3         7.000
    38.000             2        19.000        -0.375             3        10.000
    39.000             0        19.000         0.000             3        13.000
    40.000             0        19.000         0.375             3        16.000
    41.000             0        19.000         0.750             2        18.000
    42.000             0        19.000         1.125             1        19.000
    43.000             0        19.000         1.500             1        20.000
    44.000             0        19.000         1.875             0        20.000
    45.000             0        19.000         2.250             0        20.000
    46.000             1        20.000         2.625             0        20.000
 Kolmogorov Probability = 0.6993741996864222, Max Dist = 0.25
```

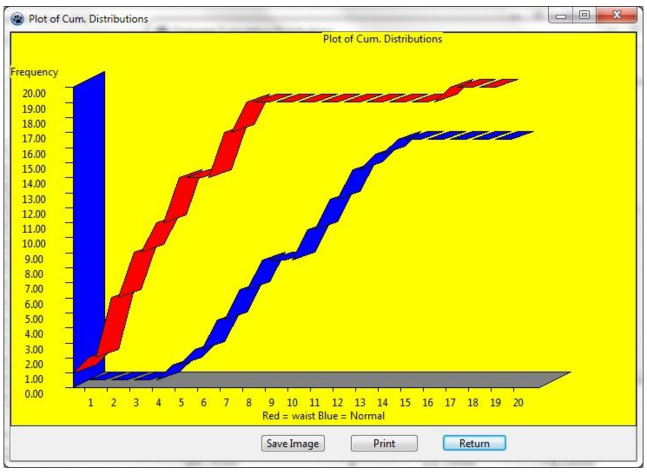


Fig. 8.19 Plot of Distributions in the Kolmogorov-Smirnov Test

Kappa and Weighted Kappa

Unweigthed and weighted Kappa Coefficents: This procedure provides both the unweighted and weighted Kappa coefficients for assessing the consistency of judgements of two raters. It also provides other measures of independence of the ratings. If nominal categories are used in the ratings, the unweighted statistic is appropriate. If the categories repesent ordinal data, the weighted Kappa statistic may be appropriate. The file labeled KappaTest4.LAZ has been used to illustrate this procedure. Shown below is the dialog form and the analysis of the data:

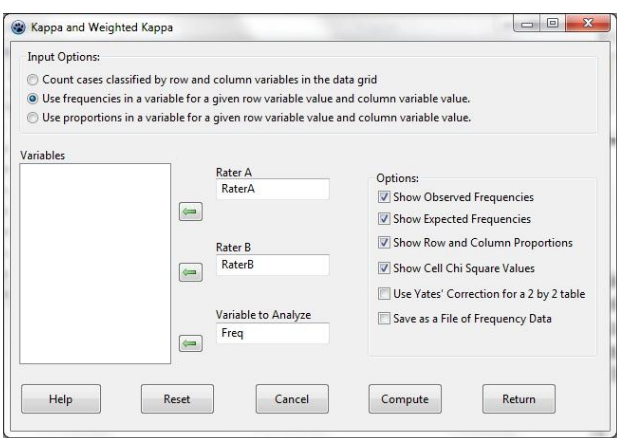


Fig. 8.20 Kappa Coefficient of Rater Agreement Form

Chi-square Analysis Results for RaterA and RaterB
No. of Cases = 100

OBSERVED FREQUENCIES

Frequencies

Variables	COL. 1	COL. 2	COL. 3	Total
Row 1	44	5	1	50
Row 2	7	20	3	30
Row 3	9	5	6	20
Total	60	30	10	100

EXPECTED FREQUENCIES with 9 cases.

Variables	COL. 1	COL. 2	COL. 3
Row 1	30.000	15.000	5.000
Row 2	18.000	9.000	3.000
Row 3	12.000	6.000	2.000

ROW PROPORTIONS with 9 cases.

Variables	COL. 1	COL. 2	COL. 3	Total
Row 1	0.880	0.100	0.020	1.000
Row 2	0.233	0.667	0.100	1.000
Row 3	0.450	0.250	0.300	1.000
Total	0.600	0.300	0.100	1.000

COLUMN PROPORTIONS with 9 cases.

Variables	COL. 1	COL. 2	COL. 3	Total
Row 1	0.733	0.167	0.100	0.500
Row 2	0.117	0.667	0.300	0.300
Row 3	0.150	0.167	0.600	0.200
Total	1.000	1.000	1.000	1.000

PROPORTIONS OF TOTAL N with 9 cases.

Variables	COL. 1	COL. 2	COL. 3	Total
Row 1	0.440	0.050	0.010	0.500
Row 2	0.070	0.200	0.030	0.300
Row 3	0.090	0.050	0.060	0.200
Total	0.600	0.300	0.100	1.000

CHI-SQUARED VALUE FOR CELLS with 9 cases.

Variables	COL. 1	COL. 2	COL. 3

```
     Row 1       6.533        6.667        3.200
     Row 2       6.722       13.444        0.000
     Row 3       0.750        0.167        8.000



 Chi-square =   45.483 with D.F. = 4. Prob. > value =    0.000

 Liklihood Ratio =   44.398 with prob. > value = 0.0000

 phi correlation = 0.6744

 Pearson Correlation r = 0.4772

 Mantel-Haenszel Test of Linear Association =   22.541 with probability > value = 0.0000

 The coefficient of contingency =    0.559

 Cramers V =    0.477

 Unweighted Kappa =   0.4915

 Observed Linear Weights with    9 cases.


 Variables
                  COL. 1       COL. 2       COL. 3
     Row 1       1.000        0.500        0.000
     Row 2       0.500        1.000        0.500
     Row 3       0.000        0.500        1.000



 Observed Quadratic Weights with    9 cases.


 Variables
                  COL. 1       COL. 2       COL. 3
     Row 1       1.000        0.750        0.000
     Row 2       0.750        1.000        0.750
     Row 3       0.000        0.750        1.000



 Linear Weighted Kappa =   0.4737
 Quadratic Weighted Kappa =   0.4545
```

Generalized Kappa

Generalized Kappa: This procedure calculates the Kappa Coefficient for objects or subjects classified into two or more categories by a group of judges or procedures. Each object is coded with a sequential integer ranging from 1 to the number of objects. Each judge is coded with an integer from 1 to the number of judges.

The categories into which the judges place the objects are coded with an integer from 1 to the number of categories. The codes for the objects, judges and category placements are column variables. The file labeled KappaTest3.LAZ has been used to demonstrate this analysis:

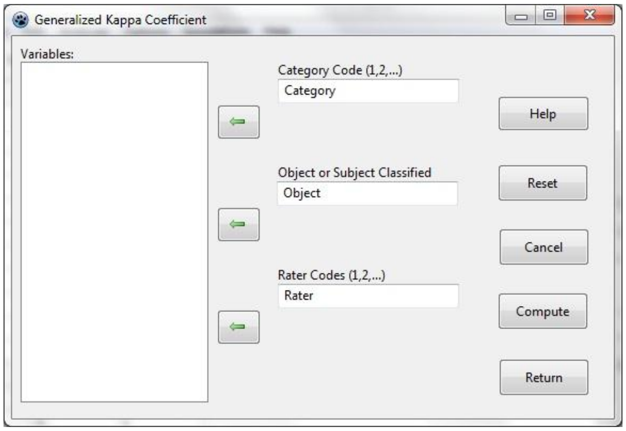


Fig. 8.21 Generalized Kappa Form

```
Generalized Kappa Coefficient Procedure
adapted from the program written by Giovanni Flammia
copywritten 1995, M.I.T. Lab. for Computer Science

2 Raters using 3 Categories to rate 1 Objects

Frequency[1][1] = 5.000000
Frequency[1][2] = 3.000000
Frequency[1][3] = 2.000000
Frequency[2][1] = 6.000000
Frequency[2][2] = 3.000000
Frequency[2][3] = 1.000000
Average_Frequency[1] = 5.500000
Average_Frequency[2] = 3.000000
Average_Frequency[3] = 1.500000
PChance = 0.215789
PObs = 0.384211
Kappa = 0.214765
z for Kappa =    0.216 with probability >    0.415
```

RIDIT Analysis

Ridit analysis was proposed by Bross4 for both the description of differences between groups on an ordered categorical scale, and the testing of the significance of those differences. The term ridit is derived from the initials of "relative to an identified distribution." The analysis begins with the identification of a population to serve as a standard or reference group. For the reference group, we estimate that the proportion of all cases with a value on the underlying continuum is falling at or below the midpoint of each interval, that is, each interval's ridit. The final values are the ridits associated with the various categories. The ridit for a category, then, is nothing but the proportion of all subjects from the reference group falling in the lower ranking categories, plus half the proportion falling in the given category.

Given the distribution of any other group over the same categories, the mean ridit for that group may be calculated. The resulting mean value is interpretable as a probability. The mean ridit for a group is the probability that a randomly-selected case from it will get better score than a randomly-selected case from the standard group. Mathematically, the mean ridit for the reference group must always be .5. This is consistent with the fact that, if two cases are randomly selected from the same population, the first case will be at least as high half the time, and will be at least as low also half the time.

Pairwise comparisons. -In most clinical studies the most sensible comparisons are those pairwise contrasts comparing one treatment group with another. There are, in general, K = k(k+l)/2 possible pairwise comparisons among the k+l groups. Critical ratio tests are presented for comparing each group with the standard and each group with the others. As a control for the increased likelihood of falsely finding significance merely because several tests were performed, we recommend the Bonferroni criterion. If the desired overall significance level is α, each comparison should be tested at the significance level α/K. Thus, if α = 0.05 and the number of groups is K = 6, α/K = 0.0083 and the corresponding critical normal curve value is 2.64. This is the criterion used for adjudging the significance of each individual pairwise comparison.

Confidence intervals. - The standard errors defined explicitly or implicitly may be used to set confidence limits about the probability that a typical case in one group obtains a higher score than a typical case in another. In order to assure that the overall confidence in the entire set of intervals is at least 100(1 - α%) (usually 95%), the Bonferroni constant, say B, should be the factor multiplying the standard error, and not the usual 1.96.

An Example. A file labeled "TEETH.LAZ" contains results from a dental study of pain suffered by patients using four different pain relief treatments. The subjects indicated degree of pain felt after a given period of time following the dental work.

The dialog form used in the analysis of this data is shown below:

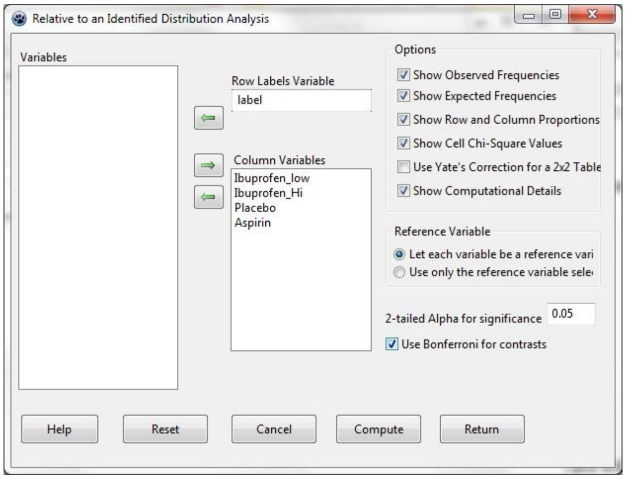


Fig. 8.22   RIDIT Analysis Dialogue Form

When the Compute button was clicked, the following results were obtained:

Chi-square Analysis Results
No. of Cases = 365

OBSERVED FREQUENCIES

Frequencies

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin	Total
None	0	1	0	1	2
Poor	6	3	18	4	31
Fair	10	5	10	11	36
Good	17	25	37	25	104
Excellent	61	52	32	47	192
Total	94	86	97	88	365

EXPECTED FREQUENCIES with   5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.515	0.471	0.532	0.482
Poor	7.984	7.304	8.238	7.474
Fair	9.271	8.482	9.567	8.679
Good	26.784	24.504	27.638	25.074
Excellent	49.447	45.238	51.025	46.290

ROW PROPORTIONS with   5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin	Total
None	0.000	0.500	0.000	0.500	1.000
Poor	0.194	0.097	0.581	0.129	1.000
Fair	0.278	0.139	0.278	0.306	1.000
Good	0.163	0.240	0.356	0.240	1.000
Excellent	0.318	0.271	0.167	0.245	1.000
Total	0.258	0.236	0.266	0.241	1.000

COLUMN PROPORTIONS with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin	Total
None	0.000	0.012	0.000	0.011	0.005
Poor	0.064	0.035	0.186	0.045	0.085
Fair	0.106	0.058	0.103	0.125	0.099
Good	0.181	0.291	0.381	0.284	0.285
Excellent	0.649	0.605	0.330	0.534	0.526
Total	1.000	1.000	1.000	1.000	1.000

CHI-SQUARED VALUE FOR CELLS with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.515	0.593	0.532	0.556
Poor	0.493	2.536	11.567	1.615
Fair	0.057	1.430	0.020	0.620
Good	3.574	0.010	3.171	0.000
Excellent	2.700	1.011	7.093	0.011

Chi-square = 38.103 with D.F. = 12. Prob. > value = 0.000

Liklihood Ratio = 38.318 with prob. > value = 0.0001

phi correlation = 0.3231

Pearson Correlation r = -0.1158

Mantel-Haenszel Test of Linear Association = 4.884 with probability > value = 0.0271

The coefficient of contingency = 0.307

Cramers V = 0.187
ANALYSIS FOR STANDARD Ibuprofen_low

Frequencies Observed

Frequencies

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0	1	0	1
Poor	6	3	18	4
Fair	10	5	10	11
Good	17	25	37	25
Excellent	61	52	32	47

Column Proportions Observed with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.012	0.000	0.011
Poor	0.064	0.035	0.186	0.045
Fair	0.106	0.058	0.103	0.125
Good	0.181	0.291	0.381	0.284
Excellent	0.649	0.605	0.330	0.534

Ridit calculations for Ibuprofen_low with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.064	0.032	0.000	0.032
Fair	0.106	0.053	0.064	0.117
Good	0.181	0.090	0.170	0.261
Excellent	0.649	0.324	0.351	0.676

Ridit calculations for Ibuprofen_Hi with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.012	0.006	0.000	0.006
Poor	0.035	0.017	0.012	0.029
Fair	0.058	0.029	0.047	0.076
Good	0.291	0.145	0.105	0.250
Excellent	0.605	0.302	0.395	0.698

Ridit calculations for Placebo with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.186	0.093	0.000	0.093
Fair	0.103	0.052	0.186	0.237
Good	0.381	0.191	0.289	0.479
Excellent	0.330	0.165	0.670	0.835

Ridit calculations for Aspirin with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.011	0.006	0.000	0.006
Poor	0.045	0.023	0.011	0.034
Fair	0.125	0.063	0.057	0.119
Good	0.284	0.142	0.182	0.324
Excellent	0.534	0.267	0.466	0.733

Ridits for all variables with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin

None	0.000	0.006	0.000	0.006
Poor	0.032	0.029	0.093	0.034
Fair	0.117	0.076	0.237	0.119
Good	0.261	0.250	0.479	0.324
Excellent	0.676	0.698	0.835	0.733

Mean RIDITS Using the Reference Values with 5 valid cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
	0.500	0.492	0.340	0.451

Overall mean for RIDITS in non-reference groups := 0.4244
Chisquared := 27.695 with probability < 0.0000

z critical ratios with 5 valid cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
	0.000	-0.182	-3.823	-1.146

significance level used for comparisons := 2.394
Ibuprofen_Hi vs Ibuprofen_low not significant
Placebo vs Ibuprofen_low significant
Aspirin vs Ibuprofen_low not significant
ANALYSIS FOR STANDARD Ibuprofen_Hi

Frequencies Observed

Frequencies

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0	1	0	1
Poor	6	3	18	4
Fair	10	5	10	11
Good	17	25	37	25
Excellent	61	52	32	47

Column Proportions Observed with 5 cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.012	0.000	0.011
Poor	0.064	0.035	0.186	0.045
Fair	0.106	0.058	0.103	0.125
Good	0.181	0.291	0.381	0.284
Excellent	0.649	0.605	0.330	0.534

Ridit calculations for Ibuprofen_low with 5 cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.064	0.032	0.000	0.032
Fair	0.106	0.053	0.064	0.117
Good	0.181	0.090	0.170	0.261
Excellent	0.649	0.324	0.351	0.676

Ridit calculations for Ibuprofen_Hi with 5 cases.

```
Variables
          Ibuprofen_low Ibuprofen_Hi      Placebo      Aspirin
      None       0.012        0.006        0.000        0.006
      Poor       0.035        0.017        0.012        0.029
      Fair       0.058        0.029        0.047        0.076
      Good       0.291        0.145        0.105        0.250
 Excellent       0.605        0.302        0.395        0.698



Ridit calculations for Placebo with    5 cases.


Variables
          Ibuprofen_low Ibuprofen_Hi      Placebo      Aspirin
      None       0.000        0.000        0.000        0.000
      Poor       0.186        0.093        0.000        0.093
      Fair       0.103        0.052        0.186        0.237
      Good       0.381        0.191        0.289        0.479
 Excellent       0.330        0.165        0.670        0.835



Ridit calculations for Aspirin with    5 cases.


Variables
          Ibuprofen_low Ibuprofen_Hi      Placebo      Aspirin
      None       0.011        0.006        0.000        0.006
      Poor       0.045        0.023        0.011        0.034
      Fair       0.125        0.063        0.057        0.119
      Good       0.284        0.142        0.182        0.324
 Excellent       0.534        0.267        0.466        0.733



Ridits for all variables with    5 cases.


Variables
          Ibuprofen_low Ibuprofen_Hi      Placebo      Aspirin
      None       0.000        0.006        0.000        0.006
      Poor       0.032        0.029        0.093        0.034
      Fair       0.117        0.076        0.237        0.119
      Good       0.261        0.250        0.479        0.324
 Excellent       0.676        0.698        0.835        0.733



Mean RIDITS Using the Reference Values with    5 valid cases.

VariablesIbuprofen_low Ibuprofen_Hi      Placebo      Aspirin
                 0.508        0.500        0.339        0.454

Overall mean for RIDITS in non-reference groups :=   0.4322
Chisquared :=   28.613 with probability <   0.0000

z critical ratios with    5 valid cases.

VariablesIbuprofen_low Ibuprofen_Hi      Placebo      Aspirin
                 0.182        0.000       -3.772       -1.040

significance level used for comparisons :=    2.394
Ibuprofen_low vs Ibuprofen_Hi not significant
Placebo vs Ibuprofen_Hi significant
Aspirin vs Ibuprofen_Hi not significant
ANALYSIS FOR STANDARD Placebo
```

Frequencies Observed

Frequencies

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0	1	0	1
Poor	6	3	18	4
Fair	10	5	10	11
Good	17	25	37	25
Excellent	61	52	32	47

Column Proportions Observed with 5 cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.012	0.000	0.011
Poor	0.064	0.035	0.186	0.045
Fair	0.106	0.058	0.103	0.125
Good	0.181	0.291	0.381	0.284
Excellent	0.649	0.605	0.330	0.534

Ridit calculations for Ibuprofen_low with 5 cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.064	0.032	0.000	0.032
Fair	0.106	0.053	0.064	0.117
Good	0.181	0.090	0.170	0.261
Excellent	0.649	0.324	0.351	0.676

Ridit calculations for Ibuprofen_Hi with 5 cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.012	0.006	0.000	0.006
Poor	0.035	0.017	0.012	0.029
Fair	0.058	0.029	0.047	0.076
Good	0.291	0.145	0.105	0.250
Excellent	0.605	0.302	0.395	0.698

Ridit calculations for Placebo with 5 cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.186	0.093	0.000	0.093
Fair	0.103	0.052	0.186	0.237
Good	0.381	0.191	0.289	0.479
Excellent	0.330	0.165	0.670	0.835

Ridit calculations for Aspirin with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.011	0.006	0.000	0.006
Poor	0.045	0.023	0.011	0.034
Fair	0.125	0.063	0.057	0.119
Good	0.284	0.142	0.182	0.324
Excellent	0.534	0.267	0.466	0.733

Ridits for all variables with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.006	0.000	0.006
Poor	0.032	0.029	0.093	0.034
Fair	0.117	0.076	0.237	0.119
Good	0.261	0.250	0.479	0.324
Excellent	0.676	0.698	0.835	0.733

Mean RIDITS Using the Reference Values with 5 valid cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
	0.660	0.661	0.500	0.616

Overall mean for RIDITS in non-reference groups := 0.6459
Chisquared := 53.073 with probability < 0.0000

z critical ratios with 5 valid cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
	3.823	3.772	0.000	2.730

significance level used for comparisons := 2.394
Ibuprofen_low vs Placebo significant
Ibuprofen_Hi vs Placebo significant
Aspirin vs Placebo significant
ANALYSIS FOR STANDARD Aspirin

Frequencies Observed

Frequencies

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0	1	0	1
Poor	6	3	18	4
Fair	10	5	10	11
Good	17	25	37	25
Excellent	61	52	32	47

Column Proportions Observed with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.012	0.000	0.011
Poor	0.064	0.035	0.186	0.045
Fair	0.106	0.058	0.103	0.125
Good	0.181	0.291	0.381	0.284
Excellent	0.649	0.605	0.330	0.534

Ridit calculations for Ibuprofen_low with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.064	0.032	0.000	0.032
Fair	0.106	0.053	0.064	0.117
Good	0.181	0.090	0.170	0.261
Excellent	0.649	0.324	0.351	0.676

Ridit calculations for Ibuprofen_Hi with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.012	0.006	0.000	0.006
Poor	0.035	0.017	0.012	0.029
Fair	0.058	0.029	0.047	0.076
Good	0.291	0.145	0.105	0.250
Excellent	0.605	0.302	0.395	0.698

Ridit calculations for Placebo with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.000	0.000	0.000
Poor	0.186	0.093	0.000	0.093
Fair	0.103	0.052	0.186	0.237
Good	0.381	0.191	0.289	0.479
Excellent	0.330	0.165	0.670	0.835

Ridit calculations for Aspirin with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.011	0.006	0.000	0.006
Poor	0.045	0.023	0.011	0.034
Fair	0.125	0.063	0.057	0.119
Good	0.284	0.142	0.182	0.324
Excellent	0.534	0.267	0.466	0.733

Ridits for all variables with 5 cases.

Variables

	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
None	0.000	0.006	0.000	0.006
Poor	0.032	0.029	0.093	0.034
Fair	0.117	0.076	0.237	0.119
Good	0.261	0.250	0.479	0.324
Excellent	0.676	0.698	0.835	0.733

Mean RIDITS Using the Reference Values with 5 valid cases.

Variables	Ibuprofen_low	Ibuprofen_Hi	Placebo	Aspirin
	0.549	0.546	0.384	0.500

```
Overall mean for RIDITS in non-reference groups :=   0.4902
Chisquared :=   20.447 with probability <   0.0001

z critical ratios with    5 valid cases.

VariablesIbuprofen_low Ibuprofen_Hi      Placebo      Aspirin
                 1.146        1.040       -2.730        0.000

significance level used for comparisons :=    2.394
Ibuprofen_low vs Aspirin not significant
Ibuprofen_Hi vs Aspirin not significant
Placebo vs Aspirin significant
```

Notice that we chose to let ***each*** group be the comparison standard. This permitted comparisons among each of the groups. Typically however one would select only one group as a standard with which to compare to the other groups. If “aspirin” was the standard comparison group, only the “placebo” treatment group was significantly different.

Scheirer-Ray-Hare Test

The SRH-Test is similar to a two-way analysis of variance but may be utilized when the dependent variable is ordinal, for example, ranks rather than in interval or ratio measurement. The sums of squared deviations are computed in the same manner as the regular anova but the tests are NOT based on the ratio with the error term. The sums of squares for the two factors and their interaction term are divided by the mean square for the total. The resulting statistic is labeled “H” and the probability is obtained using the chi-squared test with the degrees of freedom corresponding to the factor or interaction terms. As is often the case with non-parametric tests, the power to detect the alternate hypotheses is lower than for equivalent parametric tests.

The output obtained is that of the traditional parametric two-way anova but with the addition of the H and probability of a greater H. By this means you can see that the power is less than the traditional anova. As an interesting experiment, take a set of data for the traditional two-way anova, transform the dependent variable using the “transform” option under the Variables menu, and compare the analysis of the original data with the Block ANOVA procedure and the ranked data with the SRH-Test.

Median Polishing

Median polish is a technique invented by J. W. Tukey (see Tukey [1, p. 366]) for extracting row and column effects in a two-way data layout using medians rather than arithmetic means, and therefore possessing the good robustness properties held by other medianlike procedures. For a good explanation and examples, see Velleman and Hoaglin [2, Chap. 8].

METHOD Add to the two-way layout a column of row effects and a row of column effects, both initially all zeros, and a single overall effect term. In every row (including the row of column effects), subtract the row median from all entries and add the row median to the row effect. Operate similarly on columns instead of rows, then return to operate on rows, then columns, . . ., etc. The procedure terminates when the twoway layout of residuals has zero median in every row and column, and where the row and column effects each have median zero. Thus, if x_{ij} is the entry in row i, column j, if r_{ij} is the corresponding residual, μ the overall effect, α_i the ith row effect, and β_j the jth column effect, then $x_{ij} = \mu + \alpha_i + \beta_j + r_{ij}$, with $\text{median}_i(\alpha_i) = 0 = \text{median}_j(\beta_j)$, and $\text{median}_i(r_{ij}) = 0 = \text{median}_j(r_{ij})$, all i, j. This decomposition is completely analogous to that in two-way analysis of variance* (ANOVA), using medians instead of means. Note that: (i) The method is nonunique. Starting by operating on columns rather than rows may lead to a different (but qualitatively similar) answer. (ii) Instead of terminating in a finite number of steps, the iterations might converge geometrically to the solution.

Chapter 9. Statistical Process Control

Introduction

Statistical Process Control (SPC) has become a major factor in the reduction of manufacturing process errors over the past years. Sometimes known as the Demming methods for the person that introduced them to Japan and then the United States, they have become necessary tools in quality control processes. Since many of the employees in the manufacturing area have limited background in statistics, a large dependency has been built on the creation of charts and their interpretation. The statistics which underlay these charts are often those we have introduced in previous sections. The unique aspect of SPC is in the presentation of data in the charts themselves.

XBAR Chart

In quality control, observations are typically made in "lots", that is, a number of observations are made on some product's manufacturing process or the product itself at periodic intervals. For example, in the manufacture of metal bolts, the length of bolts being turned out may be sampled each hour of the day. The means and standard deviation of these sample lots may then be calculated and plotted with lines drawn to show the overall mean and upper and lower "control limits" indicating whether or not a process may be "out of control". One area of confusion which exists is the language used by industrial people in indicating their level of process control. You may hear the expression that "we employ control to 6 sigmas." They **do not mean** they use 6 standard deviations as their upper and lower control limits but rather that the probability of being out of control is that associated with the normal curve probability of a value being 6 standard deviations or greater (a very small value.) This confusion of standard deviations (sigmas) and the probability associated with departures from the mean under the normal distribution assumption is unfortunate. When you select the sigma values for control limits, the limits for 1 sigma are much closer to the mean that for 3 sigma. You may, of course, select your own limits that you feel are practical for your process control. Since variation in raw materials, tool wear, shut-down costs for replacement of worn tool parts, etc. may be beyond your control, limits must be set that maximize quality and minimize costs.

Range Chart

As tools wear the products produced may begin to vary more and more widely around the values specified for them. The mean of a sample may still be close to the specified value but the range of values observed may increase. The result is that more and more parts produced may be under or over the specified value. Therefore quality assurance personnel examine not only the mean (XBAR chart) but also the range of values in their sample lots.

S Control Chart

The sample standard deviation, like the range, is also an indicator of how much values vary in a sample. While the range reflects the difference between largest and smallest values in a sample, the standard deviation reflects the square root of the average squared distance around the mean of the values. We desire to reduce this variability in our processes so as to produce products as similar to one another as is possible. The S control chart plot the standard deviations of our sample lots and allows us to see the impact of adjustments and improvements in our manufacturing processes.

CUSUM Chart

The cumulative sum chart, unlike the previously discussed SPC charts (Shewart charts) reflects the results of all of the samples rather than single sample values. It plots the cumulative sum of deviations from the mean or nominal specified value. If a process is going out of control, the sum will progressively go more positive or negative across the samples. If there are M samples, the cumulative sum S is given as

$$S = \sum_{i=1}^{M} (\bar{X}_i - \mu_o)$$ Where $\bar{X}_i$ is the observed sample mean and μ_o is the nominal value or (overall mean.)

It is often desirable to draw some boundaries to indicate when a process is out of control. By convention we use a standardized difference to specify this value. For example with the boltsize.txt data, we might specify that we wish to be sensitive to a difference of 0.02 from the mean. To standardize this value we obtain

$$\delta = \frac{0.02}{\sigma_x}$$

or using our sample values as estimates obtain

$$\delta = \frac{0.02}{S_x} = \frac{0.02}{0.359} = 0.0557$$

A "V Mask" is then drawn starting at a distance "d" from the last cumulative sum value with an angle θ back toward the first sample deviation. In order to calculate the distance d we need to know the probabilities of a Type I and Type II error, that is, the probability □ of incorrectly concluding that a shift to out-of-control has taken place and the probability □ of failing to detect an out-of-control condition. If these values are specified then we can obtain the distance d as

$$d = \left(\frac{2}{\delta^2}\right) \ln \left(\frac{1 - \beta}{\alpha}\right)$$

When you run the CUSUM procedure you will note that the alpha and beta error rates have been set to default values of 0.05 and 0.20. This would imply that an error of the first type (concluding out-of-control when in fact it is not) is a more "expense" error than concluding that the process is in control when in fact it is not. Depending on the cost of shut-down and correction of the process versus scraping of parts out of tolerance, you may wish to adjust these default values.

The angle of the V mask is obtained by

$$\theta = \tan^{-1}\left(\frac{\alpha}{2k}\right)$$

where k is a scaling factor typically obtained as $k = 2\ \sigma_x$

The specification form for the CUSUM chart is shown below for the data file labeled boltsize.txt. We have specified our desire to detect shifts of 0.02 in the process and are using the 0.05 and 0.20 probabilities for the two types of errors.

p Chart

In some quality control processes the measure is a binomial variable indicating the presence or absence of a defect in the product. In an automated production environment, there may be continuous measurement of the product and a "tagging" of the product which is non-conforming to specifications. Due to variation in materials, tool wear, personnel operations, etc. one may expect that a certain proportion of the products will have defects. The p Chart plots the proportion of defects in samples of the same size and indicates by means of upper and lower control limits, those samples which may indicate a problem in the process.

Defect (Non-conformity) c Chart

The previous section discusses the proportion of defects in samples (p Chart.) This section examines another defect process in which there is a count of defects in a sample lot. In this chart it is assumed that the occurrence of defects are independent, that is, the occurrence of a defect in one lot is unrelated to the occurrence in another lot. It is expected that the count of defects is quite small compared to the total number of parts potentially defective. For example, in the production of light bulbs, it is expected that in a sample of 1000 bulbs, only a few would be defective. The underlying assumed distribution model for the count chart is the Poisson distribution where the mean and variance of the counts are equal.

Defects Per Unit u Chart

Like the count of defects c Chart described in the previous section, the u Chart describes the number of defects per unit. It is assumed that the number of units observed is the same for all samples. We will use the file labeled uChart.txt as our example. In this set of data, 25 observations of defects for 45 units each are recorded. The assumption is that defects are distributed as a Poisson distribution with the mean given as

$$\bar{u} = \frac{\Sigma c}{\Sigma n}$$ where c is the count of defects and n is the number of units observed.

and

$$UCL = \bar{u} + sigma \sqrt{\frac{\bar{u}}{n}} \quad \text{and } LCL = \bar{u} - sigma \sqrt{\frac{\bar{u}}{n}}$$

Chapter 10. Linear Programming

Introduction

Linear programming is a subset of a larger area of application called mathematical programming. The purpose of this area is to provide a means by which a person may find an optimal solution for a problem involving objects or processes with fixed 'costs' (e.g. money, time, resources) and one or more 'constraints' imposed on the objects. As an example, consider the situation where a manufacturer wishes to produce 100 pounds of an alloy which is 83% lead, 14% iron and 3% antimony. Assume he has at his disposal, five existing alloys with the following characteristics:

Alloy1	Alloy2	Alloy3	Alloy4	Alloy5	Characteristic
90	80	95	70	30	Lead
5	5	2	30	70	Iron
5	15	3	0	0	Antimony
\$6.13	\$7.12	\$5.85	\$4.57	\$3.96	Cost

This problem results in the following system of equations:

$$
\begin{array}{rcrcrcrcrcrcl}
X_1 & + & X_2 & + & X_3 & + & X_4 & + & X_5 & = & 100 \\
0.90X_1 & + & 0.80X_2 & + & 0.95X_3 & + & 0.70X_4 & + & 0.30X_5 & = & 83 \\
0.05X_1 & + & 0.05X_2 & + & 0.02X_3 & + & 0.30X_4 & + & 0.70X_5 & = & 14 \\
0.05X_1 & + & 0.15X_2 & + & 0.03X_3 & & & & & = & 3 \\
6.13X_1 & + & 7.12X_2 & + & 5.85X_3 & + & 4.57X_4 & + & 3.96X_5 & = & Z(\min)
\end{array}
$$

The last equation is known as the 'objective' equation. The first four are constraints. We wish to obtain the coefficients of the X objects that will provide the minimal costs and result in the desired composition of metals. We could try various combinations of the alloys to obtain the desired mixture and then calculate the price of the resulting alloy but this could take a very long time!

As another example: a dietitian is preparing a mixed diet consisting of three ingredients, food A, B and C. Food A contains 81.85 grams of protein and 13.61 grams of fat and costs 30 cents per unit. Each unit of food B contains 58.97 grams of protein and 13.61 grams of fat and costs 40 cents per unit. Food C contains 68.04 grams of protein and 4.54 grams of fat and costs 50 cents per unit. The diet being prepared must contain the at least 100 grams of protein and at the most 20 grams of fat. Also, because food C contains a compound that is important for the taste of the diet, there must be exactly 0.5 units of food C in the mix. Because food A contains a vitamin that needs to be included, there should also be a minimum of 0.1 units of food A in the diet. Food B contains a compound that may be poisonous when taken in large quantities, and the diet may contain a maximum of 0.7 units of food B. How many units of each food should be used in the diet so that all of the minimal requirements are satisfied, the maximum allowances are not violated, and we have a diet which cost is minimal? To make the problem a little bit easier, we put all the information of the problem in a tableau, which makes the formulation easier.

	Protein	Fat	Cost	Minimum	Maximum	Equal
Food A	81.65	13.60	\$0.30	0.10		
Food B	58.97	13.60	\$0.40		0.70	
Food C	68.04	4.54	\$0.50			0.50
Min.	100					
Max.		20				

The numbers in the tableau represent the number of grams of either protein and fat contained in each unit of food. For example, the 13.61 at the intersection of the row labelled "Food A" and the column labelled "Fat" means that each unit of food A contains 13.61 grams of fat.

Calculation

We must include 0.5 units of food C, which means that we include 0.5 * 4.54 = 2.27 grams of fat and 0.5 * 68.04 = 34.02 grams of protein in the diet, coming from food C. This means, that we have to get 100 - 34.02 = 65.98 grams of protein or more from Food A and B, and that we may include a maximum of 20 - 2.27 = 17.73 grams of fat from food A and B. We have to include a minimum of 0.1 units of food A in the diet, accounting for 8.17 grams of protein and 0.45 grams of fat. This means that we still have to include 65.98 - 8.17 = 57.81 grams of protein from food A and/or B, and that the maximum allowance for fat from A and/or B is now 17.73 - 0.45 = 17.28 grams. We should first look at the cheapest possibility, eg inclusion of food A for the extra required 57.81 grams of protein. If we include 57.81/81.65 = 0.708 units of food A, we have met the requirement for protein, and we have added 0.708 * 13.61 = 9.64 grams of fat, which is below the allowance of 17.28 grams which had remained. So we don't need any of the food B, which is more expensive, and which is contains less protein. The price of the diet is now $0.48. But what would we do, if food B was available at a lower price? We may or may not want to use B as an ingredient. The more interesting question is, at what price would it be interesting to use B as an ingredient insteaed of A? This could be approached by an iterative procedure, by choosing a low price for B, and see if the price for the diet would become less than the calculated price of $0.48.

Implementation in Simplex

A more sophisticated approach to these problems would be to use the Simplex method to solve the linear program. The sub-program 'Linear Programming', provided with LazStats can be used to enter the parameters for these problems in order to solve them.

Chapter11. MEASUREMENT

Evaluators base their evaluations on information. This information comes from a number of sources such as financial records, production cost estimates, sales records, state legal code books, etc. Frequently the evaluator must collect additional data using instruments that he or she alone has developed or acquired from external sources. This is often the case for the evaluation of training and educational programs, evaluation of personnel policies and their impacts, evaluation of social and psychological environments of the workplace, and the evaluation of proposed changes in the way people do business or work.

This chapter will give guidance in the development of instruments for making observations in the cognitive and affective domains of human behavior.

Test Theory

The sections presented below provide a detailed discussion of testing theory. You do not need to understand all of this theory to make appropriate use of tests in your evaluations, although it may help in avoiding some errors in decisions or selecting appropriate analytic tools. It is included for the "advanced" student of evaluation who is responsible as an "expert" in assisting other evaluators in correctly using and analyzing tests. If you are "afraid" of statistics, you may skip the formal "proofs" of the equations and focus primarily on the resulting equations.

Scales of Measurement

Measurement is the assignment of a label or number to an object or person to characterize that individual on the basis of an observed attribute. The manner in which we make our observations will determine our "scale of measurement".

Nominal Scales

Sometimes we observe an attribute in such a way that we can only classify an individual or object as possessing or not possessing the attribute. For example, the variable "gender" may be observed in such a way as to permit only labeling an individual as "male" or not male (female). The attribute of "country of origin" may lead us to classify individuals by their place of birth such as "USA", "Canada", "European", etc. The assignment of labels or names to objects based on a specific attribute is called a NOMINAL scale of measurement. We can, of course, arbitrarily select the labels to assign the observed individuals. Letters such as "A", "B", "C", etc. might be used or even numbers such as "1", "2", "3", etc. Notice**,** however that the use of numbers as labels may cause some confusion with the use of numbers to indicate a quantity of some attribute. When using a nominal scale of measurement, there is no attempt to indicate quantity. Coding males as 1 and females as 0, for example, would not indicate males are "greater" on some quantitative variable - we might just as well have assigned 1 to females and 0 to males!

Ordinal Scales of Measurement

Some attributes of individuals or objects may be observed in such a way that the individuals may be ordered, that is, arranged in a manner that indicates person "B" possesses more of the attribute than person "A", but less than person "C". For example, the number of correctly answered items on a test may permit us to say that John has a higher score than Mary but a lower score than Jim. (NOTE! We carefully avoided saying that John knows more than Mary but knows less than Jim. Such statements imply a direct relationship between the amount of knowledge of a subject and the number of items passed. This is virtually never the case!) When we assign numbers that only indicate the ordering of individuals on some attribute, our scale of measurement is called an ordinal scale.

We will add that comparing the means of groups measured with an ordinal scale leads to serious problems of interpretation. The median, on the other hand, is more interpretable.

Interval Scales of Measurement

There is a class of measurements known as interval scales of measurement. These refer to observing an attribute of individuals in such a way that the numbers assigned to individuals denote the relative amount of the attribute possessed by that individual in comparison to some "standard" or referent. The assignment of numbers in this way would permit a transformation (such as multiplying all numbers by a constant) that would preserve the proportional distance among the individuals. The numbers assigned do not indicate the absolute amount of the attribute - only the amount relative to the standard. For example, we might say that the average number of questions answered correctly on a test of 100 items measuring recall of nonsense words by a very large population of 18 year old males constitutes our "standard". IF all items are equally difficult to recall, we might use the proportion of the standard number of items recalled as an interval measure of recall ability. That is, the difference between Mary who obtains a score of 20 and John who receives a score of 40 is proportional to the difference between John and Jim who receives a score of 60. Even if we multiply their scores by 100, the distance between Mary, John and Jim is proportionally the same! Again note that the proportion of the standard number of items correctly recalled is NOT a measure of individual's ability to recall items in general. It is only their ability to recall the carefully selected items of this test in comparison to the standard that is measured. A different set of items could lead to assignment of a completely different set of numbers to each individual with different relative distances among the individuals. As another example, consider a measure of individual "wealth". Assume wealth is defined as the total of a persons debts and credits using the standard "dollar". We may clearly have individuals with negative "wealth" (debts exceed credits) and individuals with "positive" wealth (credits exceed debts). Our wealth scale has equal intervals (dollars). We can make statements such as John has 20 dollars more wealth than Mary but five dollars less wealth than Jim. In other words, we can represent the distance among our individuals as well as their order. Note, however, that an individual with a wealth score of zero (debts = assets) is NOT broke, that is, have an absence of wealth. With an interval scale of measurement 0.0 does NOT mean an absence of the attribute - only a relative amount compared to the "standard". Zero is an arbitrary point on our scale of measurement:

```
              Personal Wealth

  "Mary"          "John"          "JIM"
   ____________________________________
   -10         0         +10       +15
              dollars
```

If a test of, say, 20 history items consists of items that are equally increasing in difficulty, we may use such a test to indicate the distance among subjects administered the test. We do, however, require that if an individual misses an item with known difficulty dj, that the same individual will miss all items of greater difficulty! Please note that missing all items does not mean an absence of knowledge! (We might have included easier items.) We may also have assigned "scores" to our subjects as X = the number of items "passed" - the number of items "missed". Again, the zero point on our scale is arbitrary and does not reflect an actual amount of knowledge or absence of knowledge! Tests of intelligence, achievement or aptitude may be constructed that utilize an interval scale of measurement. Like the value of a "dollar", the "difficulty" of each item must be clearly defined. We can say, for example, that $100.00 buys an ounce of silver. We might similarly define an item of difficulty 1.0 as that item which is correctly answered by 50% of 18 year old male freshmen college students residing in the USA in 1988.

Ratio Scales of Measurement

We may sometimes observe an attribute of an individual or object in such a way that the numeric values assigned the individuals indicate the actual amount of the attribute.

For example, we might measure the time delay between the occurrence of a stimulus (e.g. the flash of a strobe light) and the observation on the surface of the brain of a change in electrical potential representing response to the stimulus. Such an observed latency may theoretically vary from 0 to infinity in whatever units of time (e.g. microseconds) that we wish to utilize. We could then make statements such as John's latency is twice as long as Mary's latency but half as long as Jim's latency. Note that a zero latency is meaningful and not an arbitrary point on the scale! Another example of a ratio scale of measurement is the distance, perhaps in inches, that a person can jump. In each case, the ratio scale of measurement has a "true" zero point on the scale which can be interpreted as an absence of the attribute. In addition, the ratio scale permits forming meaningful ratios of subject's scores. For example we might say that John can jump twice as far as Mary but Jim (who is in a wheelchair) can not jump at all! Could we ever construct a test of intelligence that yielded ratio scale numbers? What would a statement that Mary is twice as intelligent as John but half as intelligent as Jim mean? What would a score of zero intelligence mean? What would a score of 1.0 mean? Clearly, it is difficult, if not nearly impossible to construct ratio scale measures for attributes that we cannot directly observe and for which we have no meaningful "standard" with which to relate. We may, in fact, be hard-pressed to provide evidence that our psychological and educational measurement scales are even interval scales. Many are clearly only ordinal measures at best.

Reliability, Validity and Precision of Measurement

Reliability

If we stepped on and off our weight scale and each time received a different reading for our weight, we would probably go out and buy a new scale! We would say we want a reliable scale - one that consistently yields the same weight for the same object measured. When we refer to tests, the ability of a test to produce the same values when used to measure the same subjects is also called the reliability of the test. If we carefully examine the "markings" on our weight scale however, we might be surprised that there are, in fact, some variations in the values we could record. Sometimes I might weight 150.3 and the next time I get on the scale I observe 150.2. Did the scale actually give different values or was I only able to interpret the distance between the marks for 150 and 151 approximately and therefore introduce some "error" or variation in the values recorded? This lack of sufficient "in-between" markings on our scale is referred to as the precision of our measurement. If the scale is only marked in whole pounds, my precision of observation is limited to whole pounds. In fact, when the scale appears right in between 150 and 151, is the closest value 150 or 151? My error of precision is potentially 1 pound. Note that precision is NOT the same as reliability. When we speak of reliability, we are speaking of variations in repeated observations that are larger than those due to the precision of measurement alone.

In describing the reliability of an instrument, it is advantageous to have an index which describes the degree of reliability of the instrument. One popular index of reliability is the product-moment correlation between two applications of the measurement instrument to a group of individuals. For example, I might administer a history test to a group of students at 10:00 A.M. and again at 2:00 P.M. Assuming the students did not talk with each other about the test, study history during the intervening time, forget relevant history material during those four hours, etc., then the correlation between their 10:00 A.M. and 2:00 P.M. scores would estimate the reliability of the test. Our index of reliability can vary between zero (no reliability) to 1.0 (perfect reliability). Note that a reliability of less than zero is nonsense - a test cannot theoretically be less than completely unreliable!

We may also express this index of reliability as the ratio of "True Score" variance to "Observed Score Variance", that is S_t^2 / S_x^2 . We will denote this ratio as r_{xx}. This choice of r_{xx} is not capricious - we use the symbol for correlation to indicate that reliability is estimated by a product-moment correlation coefficient. The xx subscript denotes a correlation of a measure with itself. Each observed score (X) for an individual may be assumed to consist of two parts, a TRUE score (T) and an ERROR (E) score, i.e., $X_i = T_i + E_i$. For N individuals, the variance of the observed scores is

$$S_x^2 = \frac{\sum_{i=1}^{N} [(T_i + E_i) - \overline{(T_i + E_i)}]^2}{(N-1)}$$

or

$$S_x^2 = \frac{\sum_{i=1}^{N} [T_i + E_i - \bar{T} - \bar{E}]^2}{(N-1)}$$

or

$$S_x^2 = \frac{\sum_{i=1}^{N} [(T_i - \bar{T}_i) + (E_i - \bar{E})]^2}{(N-1)} \qquad (11.1)$$

If we assume that error scores (E) are normally and randomly distributed with a mean of zero and, since they are random, uncorrelated with other scores, then

$$S_x^2 = \frac{\sum_{i=1}^{N} [(T_i - \bar{T}) + E_i]^2}{(N-1)}$$

$$= \frac{\sum_{i=1}^{N}[(T_i - \bar{T})^2 + E_i^2 + (E_i T_i - E_i \bar{T})]}{(N-1)} \qquad (11.2)$$

$$= \frac{\sum_{i=1}^{N} (T_i - \bar{T})^2}{(N-1)} + \frac{\sum_{i=1}^{N} E_i^2}{(N-1)} + \frac{\sum_{i=1}^{N} E_i (T_i - \bar{T})}{(N-1)}$$

$$= S_t^2 + S_e^2 + Cov_{te} / (N-1)$$

$$= S_t^2 + S_e^2 + Cov_{te} / (N-1) * (S_t S_e)/(S_t S_e)$$

$$= S_t^2 + S_e^2 + r_{te} S_t Se$$

$= S_t^2 + S_e^2$ since the correlation of errors with true scores is zero.

Reliability is defined as

$$r_{xx} = \frac{S_t^2}{S_x^2} = \frac{S_x^2 - S_e^2}{S_x^2} = \frac{1 - S_e^2}{S_x^2} \qquad (11.3)$$

Because we cannot directly observe true scores, we must estimate them (or the variance of error scores) by some method. A variety of methods have been developed to estimate the reliability of a test. We will describe, in this unit, the one known as the Kuder-Richardson Formula 20 estimate. Other methods include the test-retest method, the corrected split-half method, the Cronbach Alpha method, etc.

The Kuder - Richardson Formula 20 Reliability

The K-R formula is based on the correlation between a test composed of K observed items and a theoretical (unobserved) parallel test of k items parallel to those of the observed test. A parallel test or item is one which yields the same means, standard deviations and intercorrelations as the original ones.

To develop the K-R 20 formula, we will begin with the correlation between two tests composed of K and k items respectively where K = k. The correlation between the total scores correct on each test is represented by

$r_{I,II}$ where

Test I scores = the sum of item scores $X_1 + X_2 +..+ X_K$
and Test II scores= the sum of item scores $x_1 = x_2 +..+ x_k$

We may therefore write the correlation as

$$r_{I,II} = r_{(X_1+X_2+..+X_K),(x_1+x_2+..+x_k)}$$

$$= \frac{\sum_{i=1}^{N} [(X_1 +..+X_K)-(\overline{X_1 +..+X_K})][(x_1 +..+x_k)-(\overline{x_1 +..x_k})]}{N \sum_{G=1}^{K} S_g^2 + N \sum_{G=1}^{K} \sum_{g=1}^{K} r_{G,g} S_G S_g} \qquad g \neq G \qquad 11.4)$$

The numerator of the above equation is the deviation cross-products of the total scores I and II. The denominator represents the variance of the composite score I. Since parallel tests have the same variance, we are assuming that the variance of test I equals that of test II. For that reason, the variance of the composite test I or II can be expressed as the sum of individual item variances plus the covariance among the items. The numerator of our correlation can be similarly expressed, that is

$$r_{I,II} = \frac{N \sum_{G=1}^{K} r_{G,G} S_G S_g + N \sum_{G=1}^{K} \sum_{g=1}^{K} r_{g,G} S_G S_g}{N \sum_{g=1}^{K} S_g^2 + N \sum_{g=1}^{K} \sum_{G=1}^{K} r_{g,G} S_g S_G} \qquad g \neq G$$

which can be further reduce as follows:

$$r_{I,II} = \frac{\sum_{g=1}^{K} r_{g,g} S_g^2 + \sum_{g=1}^{K} \sum_{G=1}^{K} r_{g,G} S_g S_G}{\sum_{g=1}^{K} S_g^2 + \sum_{g=1}^{K} \sum_{G=1}^{K} r_{g,G} S_g S_G} \qquad g \neq G$$

$$r_{I,II} = \frac{\sum_{g=1}^{K} r_{g,g} S_g^2 - \sum_{g=1}^{K} S_g^2 + \sum_{g=1}^{K} S_g^2 + \sum_{g=1}^{K} \sum_{G=1}^{K} r_{g,G} S_g S_G}{S_x^2}$$

$$r_{I,II} = \frac{\sum_{g=1}^{K} r_{g,g} S_g^2 - \sum_{g=1}^{K} S_g^2 + S_x^2}{S_x^2} \qquad (11.5)$$

Note! $r_{g,g}$ represents the correlation between parallel test items.

In an observed test of K items we would not expect to have parallel items. We must therefore estimate the correlation (or covariance) among parallel items by the correlation among non-parallel items. That is

$$\sum_{g=1}^{K} r_{g,g} S_g^2 = \frac{\sum_{g=1}^{K} \sum_{G=1}^{K} r_{g,G} S_g S_G}{(K - 1)} \qquad g \neq G$$

Note: There are K(K-1) pairings when g is not equal to G.

Since

$$S_x^2 = \sum_{g=1}^{K} S_g^2 + \sum_{g=1}^{K} \sum_{G=1}^{K} r_{g,G} S_g S_G$$

then

$$\sum_{g=1}^{K} r_{g,g} S_g = (S_x^2 - \sum_{g=1}^{K} S_g^2) / (K-1)$$

and

$$r_{I,II} = \frac{\dfrac{S_x^2 - \sum_{g=1}^{K} S_g^2}{K - 1} - \sum_{g=1}^{K} S_g^2 + S_x^2}{S_x^2}$$

$$= \frac{S_x^2 - \sum_{g=1}^{K} S_g^2}{(K-1)\, S_x^2} - \frac{\sum_{g=1}^{K} Sg2}{S_x^2} + 1$$

$$= \frac{1}{K-1} - \frac{\dfrac{\sum_{g=1}^{K} S_g^2}{S_x^2}}{K-1} - \frac{(K-1)\sum_{g=1}^{K} S_g^2}{(K-1)S_x^2} + \frac{K-1}{K-1}$$

$$= \left[\frac{1}{K-1}\right]\left[1 - \frac{\sum_{g=1}^{K} S_g^2}{S_x^2} - \frac{K\sum_{g=1}^{K} S_g^2}{S_x^2} + \frac{\sum_{g=1}^{K} S_g^2}{S_x^2} + K-1\right]$$

$$= \frac{1}{K-1}\left[K - \frac{K\sum_{g=1}^{K} S_g^2}{S_x^2}\right]$$

or $r_{I,II} = \left[\frac{K}{K-1}\right]\left[1 - \sum_{g=1}^{K} S_g^2 / S_x^2\right]$ KR#20 Formula   11.6)

We have thus derived the Kuder-Richardson Formula 20 estimate of the correlation between an observed test of K items and a theoretically parallel test of k items. Besides knowing the number of items K, one must calculate the sum of the item variances for item g = 1 to K and the total variance of the test (Sx2). We really only had to make one assumption other than the parallel test assumptions: that the covariance among UNLIKE items is a reasonable estimate of covariance among PARALLEL items.

If we might also assume that all items are equally difficult (they would have the same means and variances) then the above formula may be even further simplified to

$$r_{xx} = \frac{K}{K-1}\left[1 - \frac{\bar{X} - \bar{X}^2 / K}{S_x^2}\right] \qquad (11.7)$$

We note that in the KR#20 formula, that as the number of items K grows large, the ratio of K / (K-1) approaches 1.0 and the reliability approaches

$$r_{xx} = \frac{S_x^2 - \sum_{g=1}^{K} S_g^2}{S_x^2}$$

$$= S_t^2 / S_x^2$$

We now have an expression for the variance of true scores, that is $S_t^2 = S_x^2\, r_{xx}$. Similarly, we may obtain an expression for the variance of errors by

$$r_{xx} = (S_x^2 - S_e^2) / S_x^2$$

$$= 1.0 - S_e^2 / S_x^2$$

$$\text{or } S_e^2 = S_x^2 (1 - r_{xx}) \tag{11.8}$$

The Standard Error of Measurement, the positive root of the variance of errors is obtained as

$$S_e = S_x \sqrt{(1.0 - r_{xx})} \tag{11.9}$$

If the errors of measurement may be assumed to be normally distributed, the standard error indicates the amount of score variability to be expected with repeated measures of the same object. For example, a test that has a standard deviation of 15 and a reliability of .91 (as estimated by the KR#20 formula) would have a standard error of measurement of 15 * .3 = 4.5 . Since one standard deviation of the normal curve encompasses approximately 68.2% of the scores, we may say that approximately 68% of an individual's repeated measurements would be expected to fall within + or - 4.5 raw score points. We take note of the fact that this is the error of measurement expected of all individuals measured by a hypothetical instrument no matter what the original score level observed is. If you read about the Rasch method of test analysis, you will find that there are different estimates of measurement error for subjects with varying score levels by that method!

Validity

When we develop an instrument to observe some attribute of objects or persons, we assume the resulting scores will, in fact, relate to that attribute. Unfortunately, this is not always the case. For example, a teacher might construct a paper and pencil test of mathematics knowledge. If a student is unable to read (perhaps blind) then the test would not be valid for that individual. In addition, if the teacher included many "word" problems, the test scores obtained for students may actually measure reading ability to a greater extent than mathematics ability! The "ideal" measurement instrument yields scores indicative of only the amount (or relative amount compared with others) of the single attribute of a subject. It is NOT a score reflecting multiple attributes.

Consider, for a moment, that whenever you wanted a measure of someone's weight, your scale gave you a combination of both their height and weight! How would you differentiate among the short fat persons and the tall thin persons since they could have identical scores? If a test score reflects both mathematics and reading ability, you cannot differentiate persons good in math but poor in reading from those poor in math but good in reading!

The degree to which a test measures what it is intended to measure is called the VALIDITY of the test. Like reliability, we may use an index that varies between 0 and 1.0 to indicate the validity of a test. Again, the Pearson product-moment correlation coefficient is the basis of the validity index.

Concurrent Validity

If there exists another test in which we have confidence of it being reasonable measure of the same attribute measured by our test, we may use the p-m correlation between our test and this "criterion" test as a measure of validity. For example, assume you are constructing a new test to measure the aptitude that students have for learning a foreign language. You might administer your test and the Modern Foreign Language Aptitude Test to the same group of subjects. The correlation between the two tests would be the validity coefficient.

Predictive Validity

Some tests are intended to be used as predictors of some future attribute. For example, the Scholastic Aptitude Test (SAT) may be useful as a predictor of future Grade Point Average earned by students in their freshman year at college. When we correlate the results of a test administered at one point in time with a criterion measured at some future time, the correlation is a measure of the predictive validity of the test.

Discriminate Validity

Some tests which purportedly measure a single attribute are, as we have said, often composite measures of multiple attributes. Ideally, an English test would correlate highly with other English tests and NOT particularly high with intelligence tests, mathematics tests, mechanical aptitude tests, etc. The degree to which the correlation with similar attribute measures differs from the correlation of our test with measures of other attributes is called the discriminate validity of a test. Often the partial correlation between two tests in which the effects of a third, supposedly less related test, has been removed, is utilized as a discriminate validity coefficient. As an example, assume that your new test of English correlates .8 with student final examination scores in an English course and correlates .5 with the Stanford-Binet test of intelligence. Also assume that the final examination scores correlate .4 with the S-B IQ scores. The partial correlation of your English Test with English final examination scores can be obtained as

$$r_{y,E.I} = \frac{r_{y,E} - r_{y,I}\, r_{E,y}}{\sqrt{(1-r_{y,I}^2)(1-r_{e,y}^{\;2})}} \qquad (11.10)$$

where

$r_{y,E.I}$ is the partial correlation between your test y and the English examination scores,
$r_{y,E}$ is the correlation of your test and the English examination scores,
$r_{y,I}$ is the correlation of your test with IQ scores, and
$r_{E,y}$ is the correlation between English examination scores and IQ scores.

The obtained value would be

$$r_{y,E.I} = [(.8 - (.5)(.4)] / \sqrt{[(1\text{-}.25)(1\text{-}.16)]}$$

$$= .6 / \sqrt{[(.75)(.84)]}$$

$$= .6 / \sqrt{(.63)} = .75$$

In other words, partialling out the effects of intelligence reduced our validity from .8 to .75.

It is sometimes distressing to discover that a carefully constructed test of a single attribute often may be found to correlate substantially with a number of other tests which supposedly measure other, unrelated attributes. In our example, we partial out only the effects of one other variable, intelligence. One can use multiple regression procedures to partial out more than one variable from a correlation.

Construct Validity

The attribute we are proposing to measure with a test is often simply a hypothetical construct, that is, some attribute we think exists but which we have had to define by simple description in our language. There is often no way to directly observe the attribute. The concept of "intelligence" is such a hypothetical construct. We describe more "intelligent" people as those who learn faster and retain their learning longer. Less "intelligent" persons seem to learn at a much slower pace and have more difficult time retaining what they have learned. With such descriptions, we may construct an "intelligence" test. As you probably well know, a number of people have, in fact, done just that! Now assume that your "intelligence" test along with that of, say, three other tests of intelligence, are all administered to the same group of subjects. We could then construct the inter-correlation matrix among these four tests and ask "is there one common underlying variable that accounts for the major portion of variance and covariance within and among these tests?" This question is often answered by determining the eigenvalues and corresponding eigenvectors of the correlation matrix. If there is one particularly larger root out of the four possible roots and if the normalized corresponding eigenvalues of that root all are large, we may argue that there is validity for the construct of intelligence (at least as defined by the four tests). This technique and others similar to it are usually called "Factor Analysis". If our test "loads" (correlates) highly with the same common factor that the other tests measuring the same attribute do, then we argue the test has construct validity. This correlation (factor loading) of our test with the other measures of the same attribute is the construct validity coefficient of our test.

Content Validity

If you were to construct a test of knowledge in a specific area, say "proficiency in statistics", then the items you elect to include in your test should stand the scrutiny of experts in the field of statistics. That is, the content of your test in terms of the items you have written should be relevant to the attribute to be measured. When constructing a test, an initial decision is made as to the purpose of the test: is the purpose to demonstrate proficiency to some specified level, or is it to measure the degree of knowledge attained as compared to others. The first type of test is often referred to as a "criterion" referenced test. The second type in a normative test. With a criterion referenced test, the test writer is usually not as concerned with measuring a "single" attribute or latent variable but rather of selecting items that demonstrate specific knowledge and skills required for doing a certain job or success in some future learning activity. The norm-referenced tests, on the other hand, usually measure the degree of some predominant attribute or "latent" (underlying) variable. In either case, the test author will typically start with a "blueprint" of the domain, i.e., a list of the relevant aspects of the attribute to be measured. This blueprint may be a two-dimensional description of both the topics included in the domain as well as the levels of complexity or difficulty to be measured by items within one aspect. Once the blueprint is constructed, it is used to guide the construction of items so that the domain is adequately sampled and represented by the test. When completed, the test may be submitted to a panel of experts who are asked to classify the items into the original blueprint, evaluate the relevance of the blueprint areas and items constructed and evaluate the adequacy of the item construction. The percent of agreement among judges on a particular item as being appropriate or not being appropriate as a measure of the attribute can be used as an indicator of content validity. The reliability of judgments across a set of items may be used to measure the consistency of the judges themselves. A large proportion of the test items should be judged satisfactory by a high percentage of the judges in order to say that the instrument has content validity.

Effects of Test Length

Tests of achievement, aptitude, and ability may vary considerably in their number of items, i.e. test length. Tests composed on positively correlated items that are longer will display higher reliability then shorter tests. The correlation of reliable measures with other variables will tend to be higher than the correlation of less reliable measurements, thus the predictive validity, concurrent validity, etc. will be higher for the longer test.

Reliability for tests that have been changed in length by a factor of K can be estimated by the Spearman-Brown "prophecy" formula:

$$R_{kk} = \frac{K\ r_{11}}{1 + (K - 1)\ r_{11}} \tag{11.11}$$

where r_{11} is the reliability of the original test,
and K is the multiplication factor for lengthening (or shortening) the test.

As an example, assume you have constructed a test of 20 items and have obtained a reliability estimate of 0.60 . You are interested in estimating the reliability of the test if you were to double the number of items with items that are similar in inter-correlations, means and variances with the original 20 items. The factor K is 2 since you are doubling the length of the test. Your estimate would be:

$$R_{kk} = \frac{(2)(0.60)}{1 + (2-1)(.60)} = 0.75$$

Therefore, doubling the length of your test would result in an estimated reliability of 0.75, a sizable increase above the original 0.60. The formula can also be used to estimate the reliability of a shortened test constructed by sampling items from a longer test. For example a test of 100 items with a reliability of 0.90 could be used to produce a 25 item short-form test. The reliability would be

$$R_{kk} = \frac{(0.25)(0.90)}{1 + (0.25 - 1)(0.90)} = 0.6923$$

Note that in this case K = 0.25 since the test length has been changed by a factor of one fourth of the original length.

The Spearman-Brown formula can also be used to estimate the effects on a validity coefficient when either the test or the criterion measure have been extended in length. First we note that if a test is extended in length indefinitely (infinite length) then the reliability approaches 1.0. This permits us to estimate the validity between two measures, either or which (or both) have been extended in length. For example, the correlation between a test that has been extended by a factor of K and another test that has been extended by a factor of L is given by :

$$R_{KL} = \frac{r_{1I}}{\sqrt{1/K + (1 - 1/K)r}\ \sqrt{1/L + (1 - 1/L)r_{II}}} \tag{11.12}$$

where r_{1I} is the correlation between the two tests, r_{11} and r_{II} are the reliabilities of the two tests and K and L are the factors for extending the two tests.

If only one of the tests, say for example test I above, is made infinitely long so that its reliability approaches 1.0, then the above formula reduces to

$$R_{1\infty} = \frac{r_{1I}}{\sqrt{r_{II}}}$$

The above formula is useful in estimating the validity of a test correlated with a criterion measured without error. In addition, we may be interested in estimating the correlation of a test and criterion both of which have been adjusted for unreliability. This would estimate the correlation between the True scores of each instrument and is given by

$$R_{\infty\infty} = \frac{r_{1I}}{\sqrt{r_{11}\ r_{II}}} \tag{11.13}$$

Composite Test Reliability

Teachers often base course grades on the basis of a combination of tests administered over the period of the semester. The teacher usually, however, desires to give different weights to the tests. For example, the teacher may wish to weight tests 1, 2 as 1/4 of the total grade and the final exam (test 3) as 1/2 of the grade. Since the tests may vary considerably in length, mean, variance and reliability, one cannot simply add the weighted raw scores achieved by each student to get a total score. Doing so would give greater weight than intended to the more variable test and less weight than intended to the less variable test. A preferable method of obtaining the total weighted score would be first to standardize each test to a common mean and standard deviation. This is usually done with the z score transformation, i.e.

$$z_i = \frac{(Xi - \bar{X})}{Sx} \tag{11.14}$$

Each subject's z score for a test may then be weighted with the desired test weight and the sum of the weighted z scores be used as the total score on which grades are based. The reliability of this composite weighted z score can be estimated by the following formula:

$$R_{ww} = \frac{WCW'}{WRW'} \tag{11.15}$$

where R_{ww} is the reliability of the composite,
W is a row vector of weights and W' is the column transpose of W,
R is the correlation matrix among the tests and
C is the R matrix with the diagonal elements replaced with estimates of the individual test reliabilities.

As an example, assume a teacher has administered three tests during a semester course and obtains the following information:

	CORRELATIONS		
TEST	1	2	3
1	1.0	.6	.4
2	.6	1.0	.5
3	.4	.5	1.0
Reliability	.7	.6	.8
Weights	.25	.25	.50

The reliability of the composite score would then be obtained as:

$$R_{ww} = \frac{(.25\ \ .25\ \ .50)\begin{vmatrix} .7 & .6 & .4 \\ .6 & .6 & .5 \\ .4 & .5 & .8 \end{vmatrix}\begin{pmatrix} .25 \\ .25 \\ .50 \end{pmatrix}}{(.25\ \ .25\ \ .50)\begin{vmatrix} 1.0 & .6 & .4 \\ .6 & 1.0 & .5 \\ .4 & .5 & 1.0 \end{vmatrix}\begin{pmatrix} .25 \\ .25 \\ .50 \end{pmatrix}}$$

$$= 0.861$$

The above equation utilizes matrix multiplication to obtain the solution. If you have not used matrix algebra before, you may need to consult an elementary text book in matrix algebra to familiarize yourself with the basic operations.

Reliability by ANOVA

Sources of Error - An Example

In the previous sections, an observed score for an individual on a test was considered to consist of two parts, true score and error score, i.e. X = T + E. Error scores were assumed to be random with a mean of zero and uncorrelated with the true score. We now wish to expand our understanding of sources of errors and introduce a method for estimating components of error, that is, analyzing total observed score variance into true score variance and one or more sources of error variance. To do this, we will consider a measurement example common in education - the rating of teacher performance.

A Hypothetical Situation

Assume that teachers in a certain school district are to be rated by one or more supervisor one or more times per year. Also assume that a rater employs one or more "items" in making a rating, for example, lesson plan rating, handling of discipline, peer relationships, parent conferences, grading practices, skill in presenting material, sensitivity to students, etc.. We will assume that the teachers are rated on each item using a scale of 1 to 10 points with 1 representing very inadequate to 10 representing very superior performance. We note that in this situation:

(1) teachers to be rated are a sample from a population of teachers,
(2) supervisors doing the rating are a sample of supervisors,
(3) items selected are a sample of possible teacher performance items,
(4) ratings performed are a sample of possible replications, and
(5) teacher performance on a specific item may vary from situation to situation due to variation in teacher mood, alertness, learning, etc. as well as due to situational variables such as class size, instructional materials, time of day, etc..

We are interested of course in obtaining ratings which accurately reflect the true competence of a teacher and the true score variability among teachers (perhaps to reward the most meritorious teacher, identify teachers needing assistance, and selection of teachers for promotion). We must recognize however, a number of possible sources of variance in our ratings - sources other than the "true" competence of the teachers and therefore error of measurement:

(a) variability in ratings due to items sampled from the population of possible items,
(b) variability in ratings due to the sample of supervisors used to do the ratings,
(c) variability in ratings due to the sample of teachers rated,
(d) interactions among items, teachers and supervisors.

Let us assume in our example that six teachers are rated by two supervisors (principal and coordinator) on each of four items. Assume the following data have been collected:

	Principal				Coordinator				Combined		
Item	1	2	3	4	1	2	3	4	Princ.	Coord.	Both
Teacher											
1	9	6	6	2	8	2	8	1	23	19	42
2	9	5	4	0	7	5	9	5	18	26	44
3	8	9	5	8	10	6	9	10	30	35	65
4	7	6	5	4	9	8	9	4	22	30	52
5	7	3	2	3	7	4	5	1	15	17	32
6	10	8	7	7	7	7	10	9	32	33	65
SUM	50	37	29	24	48	32	50	30	140	160	300

Item Sums for Principal + Coordinator
98 69 79 54

We now define the following terms to use in a three way analysis of variance:

X_{ijk} = the rating for teacher i on item j from supervisor k.

1.

$$\sum_{i=1}^{6}\sum_{j=1}^{4}\sum_{k=1}^{2}(X_{ijk})^2 = 2{,}214$$ Sum of Squares of single observations.

$$\sum_{j=1}^{4}\sum_{k=1}^{2}(X_{.jk})^2 = 12{,}014$$ Sum of Squares over teachers.

$\sum_{i=1}^{6} \sum_{k=1}^{2} (X_{i.k})^2 = 8{,}026$ Sum of Squares over items.

$\sum_{i=1}^{6} \sum_{j=1}^{4} (X_{ij.})^2 = 4{,}258$ Sum of Squares over supervisors.

$\sum_{k=1}^{2} (X_{..k})^2 = 45{,}200$ Sum of Squares over teachers and items.

$\sum_{j=1}^{4} (X_{.j.})^2 = 23{,}522$ Sum of Squares over teachers and supervisors

$\sum_{i=1}^{6} (X_{i..})^2 = 15{,}878$ Sum of Squares over items and supervisors.

$(X_{...})^2 = 90{,}000$ Square of grand sum of all observations.

Our analysis of variance table may contain the following sums of squares:

$$\text{Total Sums of Squares} = \sum_{i=1}^{6} \sum_{j=1}^{4} \sum_{k=1}^{2} (X_{ijk})^2 - \frac{\left(\sum_{i=1}^{6} \sum_{j=1}^{4} \sum_{k=1}^{2} X_{ijk}\right)^2}{(6)(4)(2)}$$

or SS_{total} = 2,214 - 90,000 / 48 = 339.00

$$\text{Teacher Sums of Squares} = \frac{\sum_{i=1}^{6} (X_{i..})^2}{(4)(2)} - \frac{\left(\sum_{i=1}^{6} \sum_{j=1}^{4} \sum_{k=1}^{2} X_{ijk}\right)^2}{(6)(4)(2)}$$

or $SS_{teachers}$ = 15,878 / 8 - 90,000 / 48 = 109.80

$$\text{Item Sums of Squares} = \frac{\sum_{j=1}^{4} (X_{.j.})^2}{(6)(2)} - \frac{\left(\sum_{i=1}^{6} \sum_{j=1}^{4} \sum_{k=1}^{2} X_{ijk}\right)^2}{(6)(4)(2)}$$

or SS_{items} = 23,522 / 12 - 90,000 / 48 = 85.2

$$\text{Supervisor Sum of Squares} = \frac{\sum_{k=2}^{2} (X_{..k})^2}{(6)(4)} - \frac{\left(\sum_{i=1}^{6} \sum_{j=1}^{4} \sum_{k=1}^{2} X_{ijk}\right)^2}{(6)(4)(2)}$$

or SS_{superv} = 45,200 / 24 - 90,000 / 48 = 8.3

$$\text{Teacher-Item Interaction} = \frac{\sum_{i=1}^{6}\sum_{j=1}^{4}(X_{ij.})^2}{2} - \frac{\left(\sum_{i=1}^{6}\sum_{j=1}^{4}\sum_{k=1}^{2}X_{ijk}\right)^2}{(6)(4)(2)}$$

$$- SS_{teachers} - SS_{items}$$

or SS_{TxI} = 4,258/2 - 90,000/48 - 109.8 - 85.2 = 59.00

$$\text{Teacher-Superv. Inter} = \frac{\sum_{i=1}^{6}\sum_{k=1}^{2}(X_{i.k})^2}{4} - \frac{\left(\sum_{i=1}^{6}\sum_{j=1}^{4}\sum_{k=1}^{2}X_{ijk}\right)^2}{(6)(4)(2)}$$

$$- SS_{teachers} - SS_{superv}$$

or SS_{TxS} = 8,026/4 - 90,000/48 - 109.8 - 8.3 = 13.4

$$\text{Item-Superv. Interact.} = \frac{\sum_{j=1}^{4}\sum_{k=1}^{2}(X_{.jk})^2}{(6)} - \frac{\left(\sum_{i=1}^{6}\sum_{j=1}^{4}\sum_{k=1}^{2}X_{ijk}\right)^2}{(6)(4)(2)}$$

$$- SS_{items} - SS_{superv}$$

or SS_{IxS} = 12,014/6 - 90,000/48 - 85.2 - 8.3 = 33.8

$$\text{Teacher-Item-Super.} = \sum_{i=1}^{6}\sum_{j=1}^{4}\sum_{k=1}^{2}(X_{ijk})^2 - \frac{\left(\sum_{i=1}^{6}\sum_{j=1}^{4}\sum_{k=1}^{2}X_{ijk}\right)^2}{(6)(4)(2)}$$

$$- SS_{teachers} - SS_{items} - SS_{superv}.$$

or $SS_{TxIxS} = SS_{total} - (SS_{teachers} + SS_{items} + SS_{superv} + SS_{TxI} + SS_{TxS} + SS_{IxS})$

= 339.0 - (109.8 + 85.2 + 59.0 + 13.4 + 33.8) = 29.5

The Analysis of Variance table may be summarized as :

SOURCE	D.F.	SS	MS
Teachers (T)	5	109.8	21.96
Items (I)	3	85.2	28.40
Supervisors (S)	1	8.3	8.30
T x I Interaction	15	59.0	3.93
T x S Interaction	5	13.4	2.68
I x S Interaction	3	33.8	11.27
T x I x S Inter.	15	29.5	1.97

We may now use each of the above mean squares to estimate population variance components in examining the reliability of the ratings. We have :

$S^2_{TxIxS} = MS_{TxIxS} = 1.97$

The second order interaction is our error (residual) term since we only have a single observation under each of the three facets (teachers, items and supervisors).

$S^2_{TxI} = .5(MS_{TxI} - MS_{TxIxS}) = .5(3.93) - 1.97) = 0.98$

This is our error variance attached to teacher interaction with items. Each mean square at a given level includes variance at a higher level of interaction. We subtract out that previously obtained portion. We also divide by the number of observations on which the term is based - in this case the teacher by item interaction is based on two supervisors.

$S^2_{TxS} = (1/4)(MS_{TxS} - MS_{TxIxS}) = .25(2.68 - 1.97) = .18$

This is our estimate of error due to interaction of teachers and supervisors (repeated over the four items).

$S^2_{IxS} = (1/6)(MS_{IxS} - MS_{TxIxS}) = (11.27 - 1.97)/6=1.55$

This is the estimated error variance for interaction of items and supervisors over the six teachers.

$S^2_{T} = [1/(4)(2)][\ MS_T - MS_{TxI} - MS_{TxS} + MS_{TxIxS}]$

$= (\ 21.96 - 3.93 - 2.68 + 1.97) / 8 = 2.16$

This is our estimate of variance due to differences among teachers - that variance we hope is large in comparison to error variance. It is our estimate of the teachers variance component of each rating by each supervisor.

$S^2_{I} = [1/(6)(2)][\ MS_I - MS_{TxI} - MS_{IxS} + MS_{TxIxS}]$

$= (28.4 - 3.93 - 11.27 + 1.97) / 12 = 1.26$

This is variance due to variability of ratings among the items or item "difficulty".

$S^2_{S} = [1/(6)(4)][MS_S - MS_{TxS} - MS_{IxS} + MS_{TxIxS}]$

$= (8.3 - 2.68 - 11.27 + 1.97) / 24 < 0$

This estimate of variability due to supervisors is less than zero hence considered negligible. While variance cannot be less than zero, our small sample of supervisors that apparently rated quite consistently led to this estimate. Estimates may, of course, fall above or below the population values.

We now turn to the question of estimating the reliability of our ratings. In previous sections the classical definition of reliability was given as

$$rxx = \frac{\sigma^2_{rue}}{\sigma^2_{true} + \sigma^2_{error}} = \frac{\sigma^2_{true}}{\sigma^2_{observed}}$$

The "true" score variance for J items rated by K supervisors is given by

$$S^2_{true} = (JK)^2 S^2_T = [(4)(2)]^2 2.16 = (64)(2.16) = 138.24$$

Our "observed score" variance is estimated by

$$S^2_{obs} = (JK)(JKS^2_t + JS^2_s + KS^2_i + JS^2_{TxS} + KS^2_{TxI}$$

$$+ S^2_{IxS} + S^2_{TxIxS})$$

$$= (4x2) [(4x2)2.16 + (4)0 + (2)1.26 + (4)0.18$$
$$+ (2)0.98 + 1.55 + 1.97)$$
$$= 208$$

and the ratio $S^2_{true} / S^2_{observed} = r_{xx} = 0.665$ is the estimate of the correlation that would be obtained between two sets of scores for a group of teachers rated on the basis of a random set of four items chosen for each teacher and rated by a random set of two supervisors for that teacher. Note our emphasis that this is a random effects model - each teacher could be rated on a sample of different items and by different supervisors!

In examining the sources of error, increasing the number of items would most likely reduce the largest error components (items and interaction of items with teachers and supervisors).

If the items used by each person doing the ratings is the same (fixed effects of items), the variance component for items disappears from the estimate of observed score variance giving

$$S^2_{observed} = (0 + .24 + 0.09 + 0.19 + 0.25) = 2.93$$

and $r_{xx} = 2.16 / 2.93 = 0.74$

Obviously, using the same test on all teachers yields a more precise estimate of the teacher competencies. If we also fix the supervisors so that all teachers are rated by the same two supervisors, then S^2_S and S^2_{IxS} disappears as sources of error variance and the observe score is given by

$$S^2 = S^2_T + S^2_{TxI} / J + S^2_{TxS} / K + S^2_{TxIxS} / JK$$

$$= 2.16 + 0.24 + 0.09 + 0.25 = 2.74$$

and $r_{xx} = 2.16 / 2.74 = 0.79$

By using the same items and supervisors, the reliability of the ratings has been increased from .66 to .79 .

We may further assume that our items are not a sample from a population of items but, in fact, constitute the universe of teacher behaviors to which we intend to generalize. In this case, S^2_{TxI} and S^2_I will both disappear from our error term. Our estimates of true and observe score therefore become:

$$S^2_{true} = S^2_T + S^2_{TxS} / J = 2.16 + 0.24 = 2.40$$

and

$$S^2_{observed} = S^2_{true} = (S^2_S/K + S^2_{TxS}/K + S^2_{IxS} / JK$$

$$+ S^2_{TxIxS} / JK) = 2.93$$

Therefore $r_{xx} = 2.4 / 2.93 = 0.82$

Finally, if we choose to consider only two specific supervisors as our universe of supervisors, then

$S^2_{true} = S^2_T + S^2_{TxS} / J + S^2_{TxS} / K = 2.16 + .24 + .09$

$= 2.49$

and $S^2_{observ} = S^2_{true} + S^2_{TxIxS} / JK = 2.49 + 0.25 = 2.74$

Therefore, $r_{xx} = 2.49 / 2.74 = 0.91$

Clearly, the degree to which one intends to generalize a test or rating procedure affects the reliability of the measurements for that purpose.

In the previous discussion we have examined multiple facets of reliability. We saw that the assumptions of sampling both test items and raters as well as subjects affected our estimate of reliability. We now will relate the above analysis with a simple ANOVA approach using the "Treatments by Subjects" analysis of variance program found in the Measurement Menu of the SAMPLE system. To illustrate its use, we will combine the two supervisor ratings from the above example and treat our data as consisting of six teachers who have been rated on four items. We assume we are using the population of "items" and the same raters on each teacher rated. Our data consists of the following:

	ITEM				SUM
TEACHER	1	2	3	4	
1	17	8	14	3	42
2	16	10	13	5	44
3	18	15	14	18	65
4	16	14	14	8	52
5	14	7	7	4	32
6	17	15	17	16	65
SUM	98	69	79	54	300

In calculating the sums of squares for the ANOVA, we first obtain the squares of individual ratings, squares of the sums for each teacher, squares of the sums for each item and the square of the sum of the item (or teacher) sums. These are:

$$\sum_{i=1}^{6} \sum_{j=1}^{4} (X_{ij})^2 = 4,258 \quad \text{Squares of single observations}$$

$$\sum_{i=1}^{6} (X_{i.})^2 = 15,878 \quad \text{Squares of teacher sums}$$

$$\sum_{j=1}^{4} (X_{.j})^2 = 23,522 \quad \text{Squares of item sums}$$

$$(X_{..})^2 = 90,000 \quad \text{Square of grand total}$$

The sum of squared deviations about the mean for the terms of our ANOVA are obtained using the above terms and computed as follows:

$SS_{total} = 4,258 - 90,000 / 24 = 508$

$SS_{teachers} = 15,878 / 4 - 90,000 / 24 = 219.50$

SS_{items} = 23,522 / 6 - 90,000 / 24 = 170.33

SS_{IxT} = SS_{total} - $SS_{teachers}$ - SS_{items} + 90,000 / 24

= 118.17

The SS_{items} and SS_{IxT} are often combined into a SS_{within} to represent the total sum of squares due to variation within subjects, i.e. the squared deviations of subject's scores about the subject means. The ANOVA summary table may look as follows:

SOURCE	D.F.	SS	MS	F
Among Teachers	5	219.50	43.90	5.57
Within Teachers	18	288.50	16.03	
Items	3	170.44	56.78	7.21
Teachers x Items	15	118.17	7.88	
Total	23	508.00		

The terms for our reliability are

S^2_{true} = $(MS_{observed} - MS_{TxI}) / N$

= (43.90 - 7.88) / 6 = 6.00

$S^2_{observed}$ = S_{true} + MS_{TxI} / N = 6.02 + 7.88 / 6

= 6.00 + 7.88 / 6 = 7.31

and the reliability is

$r_{xx} = S^2_{true} / S^2_{observed} = 6.00 / 7.31 = 0.82$

This reliability is called the adjusted average rating reliability on the printout from the program in your system. It reflects the reliability of ratings in which the error due to differences in average ratings by the judges or items has been removed. Essentially, the individual ratings are "adjusted" so that the column sums or means are equal. If a test of J dichotomously scored items are analyzed by both the Kuder-Richardson Formula 20 and the Treatment by Subjects ANOVA procedures, the KR#20 reliability will equal the reliability reported above.

One can also estimate a single item reliability by obtaining an average item reliability using

$$r_{single} = \frac{MS_T - MS_{TxI}}{MS_T + (J-1)MS_{TxI}} = \frac{43.9 - 7.88}{43.9 - (3)7.88} = 0.53$$

Again, this reliability reflects an adjustment for the "difficulty" of the items, that is, all ratings or items are made to reflect the same sum or average across the subjects rated. A similar result would be obtained by using the Spearman-Brown Prophecy formula where we estimate the reliability of a test reduced in length to a single item.

Should the user want to know what the reliability of the ratings or test is without adjustment for variability in mean ratings, then the following may be used:

For the unadjusted test reliability

$$r_{xx} = 1.0 - (MS_{within} / MS_T)$$

$$= 1.0 - (16.03 / 43.90) = 0.63$$

For the estimate of a single item reliability unadjusted for difference among item (or rating) means, the formula is

$$r_{xx} = \frac{MS_T - MS_{within}}{MS_T + (J-1)MS_{within}}$$

$$= (43.9 - 16.03) / (43.9 + (4-1)16.03)$$

$$= 0.30$$

Item and Test Analysis Procedures

Teachers typically construct their own tests to measure the achievement of students in their courses. In constructing the test, it is a good idea to begin with a test "blueprint" or table of specifications for the test. This test blueprint usually consists of a table in which the rows represent content or concept areas to be tested and the columns represent levels of thinking required such as classified by Bloom's taxonomy of cognitive skills. The cells may simply indicate the number of items to be written in each concept area at each level of thinking skill. For example, an elementary teacher might construct a blueprint for a test over arithmetic concepts for eighth grade students using something like the following:

LEVEL CONCEPT	Knowledge	Application	Synthesis	Evaluation
Addition	3	3	1	1
Subtraction	2	2	1	2
Multiplication	3	3	0	0
Division	2	2	2	1
Percentage	3	3	3	3
Exponents	2	3	1	1

In this example, the teacher would construct 47 items from the table of specifications. The items constructed may be of a variety of types such as multiple choice, matching, completion, problem solving, etc.. Once the test is constructed and administered to the students, the teacher may then evaluate various properties of the items and test. For example, the teacher may want to know how reliable the test is, how difficult each item was, how well each item differentiates between high and low scoring students, and how the test might be improved for subsequent use. This section describes several methods for analyzing tests and the items within tests.

Classical Item Analysis Methods

Item Discrimination

If a test is constructed to test one predominant domain or area of achievement or knowledge then each item of the test should correlate positively with a total score on the test. The total score on the test is usually obtained by awarding a value of 1 to a student if they get an item correct and a 0 if they miss it and summing across all items.

On a 47 item test, a student that gets all items correct would therefore have a total score of 47 while the student that missed all items would have a score of 0.

We can correlate each item with the total score obtained by the students. We may use the Pearson Product-Moment correlation formula (see the section on simple correlation and regression) to do our calculations. We note however that we are correlating a dichotomous variable (our item is scored 0 or 1) with a continuous variable (total scores vary from 0 to the number of items in the test). This type of correlation is also called a "Point-Biserial Correlation". Unfortunately, when one of the variables in the product-moment correlation is dichotomous, the correlation is affected by the proportion of scores in the dichotomous variable. If the proportion of 0 and 1 scores is about the same (50% for each), the correlation may approach 1.0. When the split of the dichotomous variable is quite disproportionate, say .2 and .8, then the correlation is restricted to much lower values. This certainly makes interpretation of the point-biserial correlation difficult. Nevertheless, a "good" test item will have positive correlations with the total score of the test. If the correlation is negative, it implies that more knowledgable students are more likely to have missed the item and less knowledgeable students likely to have gotten the item correct! Clearly, such an item is inconsistent with the measurement of the remaining items. Remember that the total score contains, as part of the total, the score of each item. For that reason, the point-biserial correlation will tend to be positive. A "corrected" point-biseral correlation can be obtained by first subtracting the individual item score from the total score before calculating the correlation between the item and total. If a test has many items, say more than 30, the correction will make little difference in the correlation. When only a few items are administered however, the correction should be applied.

The point-biserial correlation between test item and test total score is a measure of how well the item discriminates between low and high achievement students. It is a measure of item discrimination potential. Other item discrimination indices may also be used. For example, one may simply use the difference between the proportion passing the item in students ranking in the top half on the total score and the proportion passing the item among students in the bottom half of the class. Another index, the biserial correlation, may be calculated if one assumes that the dichotomously scored item is actually an imprecise measure of a continuous variable. The biserial correlaiton may be obtained using the formula:

$$r_{bis} = r_{pbis} \sqrt{p_i\, q_i / y_I} \qquad (11.16)$$

where r_{pbis} is the point-biserial correlation, p_i and q_i are the proportions passing and failing the item, and yi is the ordinate of the normal curve corresponding to the cumulative proportion pi.

Item difficulty

In classical test analysis, the difficulty of an item is indicated by the proportion of subjects passing the item. An easy item therefore has values closer to 1.0 while more difficult items have values closer to 0.0 . Since the mean of an item scored either 0 or 1 is the same as the proportion of subjects receiving scores of 1, the mean is the difficulty of the item. An ideal yardstick has markings equally spaced across the ruler. This permits its use to measure objects varying widely in length. Similarly, a test composed of items equally spaced in difficulty permits reasonable precision in measuring subjects that vary widely in their knowledge. With item difficulties known, one can select items along the continuum from 0 to 1.0 so that the revised instrument has approximately equal interval measurement. Unfortunately, the sample of subjects on which the item difficulty estimates are based must adequately represent all of the subjects for which the instrument is to be used. If another group of subjects that differs considerably in their knowledge is used to estimate the item difficulties, quite different estimates can be obtained. In other words, the item difficulty estimates obtained in classical test analysis methods are dependent on the sample from which they are obtained. It would clearly be desirable to have item parameter estimates that are invariant from group to group, that is, independent of the subjects being measured by those items.

In our discussion we have not mentioned errors of measurement for individual items. In classical test analysis procedures we must assume that each item measures with the same precision and reliability as all other items. We usually assume that errors of measurement for single items are normally distributed with a mean of zero and that these errors contribute proportionally to the error of measurement of the total test score. Hence the standard error of measurement is assumed equal for subjects scoring from 1 to 50 on a 50 item test!

The Item Analysis Program

The LazStats package includes a program for item analysis using the Classical test theory. The program provides for scoring test items that have been entered as 0's and 1's or as item choices coded as numbers or letters. If item choices are in your data file, you will be asked to enter the correct choice for each item so that the program may convert to 0 or 1 score values for each item. A set of items may consist of several independent sub-tests. If more than one sub-test exists, you will be asked to enter the sequence number of each item in the sub-tests. You may also elect to correct for guessing in obtaining total scores for each subject. Either rights-wrongs or rights - 1/4 wrongs may be elected. Finally, you may weigh the items of a test to give more or less credit in the total score to various items. An option is provided for printing the item scores and sub-score totals. You may elect one of several methods to estimate the reliability of the scored items. The sub-test means and standard deviations are computed for the total scores and for each item. In addition, the point-biserial correlation of each item with each sub-score total is obtained. Item characteristic curves are optionally printed for each item. The curves are based on the sub-score in which the item is included. The proportion of subjects at each decile on the sub-score that pass the item is plotted. If a reasonably large number of subjects are analyzed, this will typically result in an approximate "ogive" curve for each item with positive point-biserial correlations. Examination of the plots will reveal the item difficulty and discrimination characteristics for various ability score groups.

Item Response Theory

The past few decades has seen a rapid advance in the theories of psychological measurement. Among the more important contributions is the conceptualization of subject's responses to a single item. Simply stated, we assume that the probability of a subject correctly answering an item is a function of both subject and item parameters (stable characteristics). Usually the subject is considered to have one parameter - ability (or knowledge). The item, on the other hand, may have one or more parameters. Item difficulty is one parameter but item discrimination and chance-correctness are two other possible parameters to estimate. For example, a multiple choice item with five alternatives has a smaller probability of being correctly answered by guessing than a true-false type of question. Additionally, some items may differentiate among a broad range of student abilities while others discriminate only among subjects within a narrow range of abilities.

The functional relationship between the probability for correctly answering a question and the ability of subjects is usually represented by an item characteristic curve such as that depicted below. We might use total scores on the test as approximations of subject's ability parameter and plot the proportion of subjects in each score group that correctly answered the item.

```
PROPORTION
CORRECT
1.0                          *
0.9                     *
0.8                  *
0.7                 *
0.6                *
0.5              *
0.4            *|
0.3          *  |
0.2        *    |
0.1  *  *       |
0.0             |
________________________________________________________
  1  2  3  4  5  6  7  8  9 10  11  12  13  14
      ABILITY (ESTIMATED BY TOTAL TEST SCORE)
```

An individual's ability score may be obtained by averaging the probabilities for those items correctly answered and multiplying by the number of items in the test. In the Fig. above, a vertical line is drawn at the median (50 percentile). This represents the ability of subjects that have a 50-50 chance (odds) of passing the item. It also may be considered the difficulty of the item. Note that the probabilities of passing the item increase continuously as the total score (or ability) of the subjects increase. We say that the probability of passing the item is a monotonic increasing function of ability. Clearly, an item for which the probability of correctly answering the item decreased as subject abilities increased would not be a desirable item! The slope of the curve at the median denotes the "discriminating power" of the item. If the slope is steep, a small change in subject ability produces a relatively large change in the probability of correctly answering the item. A very shallow slope would imply a low ability of the item to differentiate among subjects widely varying in ability. Typically, an item with a steep slope will only have that steepness over a relatively small range of abilities. For that reason, one item is insufficient to measure abilities with precision over a wide range of abilities. One would ideally have an instrument composed of multiple items with steep (and equal) sloped characteristic curves that overlapped on the linear portions of the curves. The Fig. below might represent a four item test with items equally spaced in difficulty and equal in discrimination:

```
PROPORTION
CORRECT

1.00                      1     2     3     4
0.95                   1     2     3     4
0.9                 1     2     3     4
0.8                1     2     3     4
0.7               1     2     3     4
0.6              1     2     3     4
0.5             1     2     3     4
0.4            1     2     3     4
0.3           1     2     3     4
0.2          1     2     3     4
0.1        1     2     3     4
0.05    1     2     3     4
0.00 1     2     3     4
________________________________________________________________________
    1   2   3   4   5   6   7   8   9  10  11  12  13
                    ABILITY
```

It is apparent that items 1, 2, 3 and 4 above provide a different amount of information concerning the ability of subjects that differ in their ability. For example, item one provides little information about subjects that have total score ability greater than 8. Similarly, item 4 provides little information for subjects scoring below 5. The amount of discrimination information of an item for varying levels of ability is a function of the slope of the item line at each ability level. If we can describe the rate of change of ability at any point on an item characteristic curve, we can plot that rate of change against ability level. Such "plots" are called item information curves. A test information curve can similarly be plotted by summing the item information (rate of ability change) at each ability level. For an item of moderate difficulty and relatively steep slope, such an item information function might look like the Fig. below:

```
ITEM INFORMATION (Y AXIS) vs ABILITY PARAMETER (X AXIS)
(Rate of Change
in Ability)
1.0                      *
0.9                   *      *
0.8                 *           *
0.7               *               *
0.6              *                 *
0.5             *                   *
0.4            *                     *
0.3          *                         *
0.2        *                             *
0.1     *                                   *
0.0   *                                        *
_____________________________________________________________________________
      1   2   3   4   5   6   7   8   9  10  11  12  13  14
```

The One Parameter Logistic Model

In the classical approach to test theory, the item difficulty parameter is estimated by the proportion of subjects in some "norming" group that passes the item. Other methods may be used however, to estimate item difficulty parameters. George Rasch developed one such method. In his model, all items are assumed to have equal item characteristic slopes and little relevant chance probabilities. The probability of a subject answering an item correctly is given by the formula

$$P(X=1|b_i) = \frac{e^{D(d_j - b_i)}}{1.0 - e^{D(d_j - b_i)}} \tag{11.17}$$

where b_i is the ability of an individual,
d_j is the difficulty of item j, D is an arbitrary scaling or expansion factor, and
e is the constant 2.7182818.....(the base of the natural logarithm system).

An individual's ability b_i is estimated by the product of the expansion factor D and the natural log odds of obtaining a score X out of K items, i.e.,

$$b_i = D \log \frac{X}{K - X} \tag{11.18}$$

The above equation may also be solved for X, the subject's raw score X expected given his or her ability, that is

$$X_i = \frac{K e^{(b_i / D)}}{1 + e^{(b_i / D)}} \tag{11.19}$$

The expansion factor D is a value which reflects the variability of both item difficulties d_j and abilities b_i. When scores are approximately normally distributed, this value is frequently about 1.7.

The Rasch one-parameter logistic model assumes that all items in the test that are analyzed measure a common latent variable. Researchers sometimes will complete a factor analysis of their test to ascertain this unidimensional property prior to estimating item difficulties using the Rasch model. Items may be selected from a larger group of items that "load" predominantly on the first factor of the set of items.

The LazStats package includes a program to analyze subject responses to a set of items. The results include estimates of item difficulties in log units and their standard errors. Ability estimates in log units and errors of estimate are also obtained for subjects in each raw total score group. One cannot estimate abilities for subjects that miss all items or correctly answer all items. In addition, items that all subjects miss or get correct cannot be scaled. Such subjects or items are automatically eliminated by the program. The program will also produce item characteristic curves for each item and report the point-biserial correlation and the biserial correlation of each item with the total test score.

The Rasch method of calibrating item difficulty and subject ability has several desirable properties. One can demonstrate that the item difficulties estimated are independent of the abilities of subjects on which the estimates are based. For example, should you arbitrarily divide a large group of subjects into two groups, those who have total scores in the top half of the class and those who have scores in the bottom half of the class, then complete the Rasch analysis for each group, you will typically obtain item difficulties from each analysis that vary little from each other. This "sample-free" property of the item difficulties does not, of course, hold for item difficulties estimated by classical methods, i.e. proportion of a sample passing an item. Ability estimates of individuals are similarly "item-free". A subset of items selected from a pool of Rasch calibrated items may be used to obtain the same ability estimates of an individual that would be obtained utilizing the entire set of items (within errors of estimation). This aspect of ability estimation makes the Rasch scaled items ideal for "tailored" testing wherein a subject is sequentially given a small set of items which are optimized to give maximum precision about the estimated ability of the subject.

Estimating Parameters in the Rasch Model: Prox. Method

Item difficulties and subject abilities in the Rasch model are typically expressed in base e logarithm values. Typical values for either difficulties or abilities range between -3.0 and 3.0 somewhat analogous to the normally distributed z scores. We will work through a sample to demonstrate the calculations typically employed to estimate the item difficulties of a short test of 11 items administered to 127 individuals (See Applied Psychometrics by R.L. Thorndike, 1982, pages 98-100). In estimating the parameters, we will assume the test items involved the student in generating a response (not multiple choice or true false) so that the probability of getting the item correct by chance is zero. We will also assume that the items all have equal slopes, that is, that the change in probability of getting an item correct for a given change in student ability is equal for all items. By making these assumptions we need only solve for the difficulty of the item.

The first task in estimating our parameters is to construct a matrix of item failures for subjects in each total score group. A total score group is the group of subjects that have the same score on the test (where the total score is simply the total number of items correctly answered). Our matrix will have the total test score as columns and individual items as rows. Each element of the matrix will represent the number of students with the same total test score that failed a particular item. Our sample matrix is

	TOTAL TEST SCORE										TOTAL
ITEM	1	2	3	4	5	6	7	8	9	10	FAILED
1	10	10	10	7	7	4	2	1	0	0	51
2	10	14	14	12	17	12	5	1	0	0	85
3	10	14	11	11	7	6	3	0	0	0	62
4	1	1	0	1	0	0	0	0	0	0	3
5	10	8	9	6	6	3	1	1	0	0	44
6	10	14	14	15	21	21	12	6	2	1	116
7	10	14	11	13	19	22	8	5	0	1	103
8	10	14	8	8	12	7	1	0	1	0	61
9	10	14	14	14	20	18	11	4	0	1	106
10	10	14	14	14	19	20	9	9	1	2	112
11	9	10	4	4	5	2	0	0	0	0	34
No.in Grp.	10	14	14	15	22	23	13	9	2	5	127

We begin our estimation of the difficulty of each item by calculating the odds of any subject failing an item. Since the far right column above is the total number of subjects out of 127 that failed the items in each row, the odds of failing an item are

$$\text{odds} = \frac{\text{no. failing}}{\text{no. subjects - no. failing}} \qquad (11.20)$$

If we divided the numerator and denominator of the above ratio by the number of subjects we would obtain for any item i, the odds

$$\text{odds} = \frac{P_i}{1.0 - P_i} \tag{11.21}$$

Next, we obtain the natural logarithm of the odds of failure for each item. The mean and variance of these log odds are then obtained. Now we calculate the deviation of each item's log odds from the mean log odds of all items. To obtain the PROX. estimate of the item difficulty we multiply the deviation log odds by a constant Y. The constant Y is obtained by

$$Y^2 = \frac{1 + V / 2.89}{1 - UV / 8.35} \tag{11.22}$$

where V is the variance of the log odds of items and
W is the variance of the log odds of abilities.

Clearly, we must first obtain the variance of log odds for abilities before we can complete our PROX. estimates for items. To do this we must obtain the odds of subjects in each total score group obtaining their total score out of the total number of possible items. For subjects in each total score group the odds are

$$\text{odds} = \frac{\text{No. items passed}}{\text{No. items - No. items passed}} \tag{11.23}$$

For example, for subjects that have a total score of 1, the odds of getting such a score are 1 / (11 - 1) = 1 / 10 = .1. Note that if we divide the above numerator and denominator by the number of test items, the formula for the odds may be expressed as

$$\text{odds} = \frac{P_j}{1 - P_j}$$

We obtain the logarithm of the score odds for each subject, and like we did for items, obtain the mean and variance of the log odds for all subjects. The variance of subject's log odds is denoted as U in the "expansion" factor Y above. A similar expansion factor will be used to obtain Prox. estimates of ability and is calculated using

$$X^2 = \frac{1 + U / 2.89}{1 - UV / 8.35} \tag{11.24}$$

The Prox. values for items is now obtained by multiplying the expansion factor Y (square root of the Y^2 value above) times the deviation log odds for each item. The Prox. values for abilities is obtained by multiplying the corresponding expansion value X times the log odds for each score group. The calculations are summarized below:

ITEM	FAILED	PASSED	ODDS	LOG ODDS	DEVIATION	PROX.
1	51	76	.67	-0.3989	-0.61	-0.87
2	85	42	2.02	0.7050	0.49	0.70
3	62	65	.95	-0.0473	-0.26	-0.37
4	3	124	.02	-3.7217	-3.93	-5.62
5	44	83	.53	-0.6347	-0.84	-1.20
6	116	11	10.55	2.3557	2.15	3.08
7	103	24	4.29	1.4567	1.25	1.79

8	61	66	0.92	-0.0788	-0.29	-0.41
9	106	21	5.05	1.6189	1.41	2.02
10	112	15	7.47	2.0104	1.80	2.58
11	34	93	.37	-1.0062	-1.22	-1.75

MEAN LOG ODDS DIFFICULTY = 0.21
VARIANCE LOG ODDS DIFFICULTY = 2.709

TOTAL SCORE	PASSED	FAILED	ODDS	LOG ODDS	PROX. ABILITY
1	1	10	.10	-2.30	-3.93
2	2	9	.22	-1.50	-2.56
3	3	8	.38	-0.98	-1.71
4	4	7	.	-0.56	-0.94
5	5	6	.83	-0.18	-0.31
6	6	5	1.20	0.18	0.31
7	7	4	1.75	0.56	0.94
8	8	3	2.67	0.98	1.71
9	9	2	4.50	1.50	2.56
10	10	1	10.00	2.30	3.93

MEAN LOG ODDS ABILITY = -0.28
VARIANCE LOG ODDS ABILITY = 1.038

Y EXPANSION FACTOR = 1.4315
X EXPANSION FACTOR = 1.709

Theoretically, the number of subjects in total score group j that pass item i are estimates of the item difficulty di and the ability bj of subjects as given by

$$b_j - d_i = \log[\, p_{ij} / (n_j - p_{ij}) \,]$$

where p_{ij} is the proportion of subjects in score group j that pass item i and nj is the number of subjects in score group j. The Prox. estimates of difficulty and ability may be improved to yield closer estimates to the pij values through use of the Newton-Rhapson iterations of the maximum-likelihood fit to those observed values. This solution is based on the theory that

$$p_{ij} = \frac{e^{(b_j - d_i)}}{1 + e^{(b_j - d_i)}} \qquad (11.25)$$

It is possible, using this procedure, that values do not converge to a solution. The Rasch program included in the statistics package will complete a maximum of 25 iterations before stopping if the solution does not converge by that time.

If the Rasch model fits the data observed for a given item, the success and failure of each score group on an item should be adequately reproduced using the estimated parameters of the model. A chi-squared statistic may be obtained for each item by summing, across the score groups, the sum of two products: the odds of success times the number of failures and the odds of failure times the number of successes. This chi-squared value has degrees of freedom N - n where N is the total number of subjects and k is the total number of score groups. It should be noted that subjects with scores of 0 or all items correct are eliminated from the analysis since log odds cannot be obtained for these score groups. In addition, items which are failed or passed by all subjects cannot be scaled and are eliminated from the analysis.

Measuring Attitudes, Values, Beliefs

The evaluator of training workshops is often as interested in how participants "feel" about their training as well as how much they have learned and retained. The testing theory presented above dealt primarily with the measure of knowledge and gave the methods for defining and testing the reliability and validity of those measures. In a similar manner, we may be interested in developing and administering instruments to measure such things as:

(a) attitudes toward management
(b) attitudes toward training experiences
(c) attitudes toward protected classes (women, minorities)
(d) attitudes toward alternative work arrangements
(e) attitudes toward safety codes and/or practices
(f) attitudes toward personnel in other departments

It is generally recognized that the way people feel about each other, their work environment and their work characteristics are important to their productivity and longevity on the job. This section is devoted to helping the evaluator construct instruments to measure such attitudes.

Methods for Measuring Attitudes

Most of you have completed at least one questionnaire of the following type:

THESIS RESEARCH
SURVEY

DIRECTIONS:

Listed below are ten statements about thesis research. Please indicate whether you agree or disagree with each statement. Circle the A if you tend to agree with the statement or circle the D if you tend to disagree with the statement. Do not spend too much time thinking about each statement. Use your first impression. GO AHEAD!

A D 1. The research one does for his or her thesis may determine the line of research they pursue the rest of their life.

A D 2. The only reason theses are required is because the current faculty had to do one in order to graduate.

A D 3. Most theses make little contribution to the body of knowledge in a discipline.

A D 4. A thesis can demonstrate your ability to be creative and thorough in conducting a research project.

A D 5. Unless you almost have a major in statistics, its very difficult to complete a useful thesis.

A D 6. Reading a thesis is right up there with reading a telephone book for pleasure.

A D 7. Certain fields like clinical psychology, business and technology where the graduate is not going to be a college professor should not require a thesis.

A D 8. Ten years after completing their degree, most students are ashamed of their thesis.

A D 9. The whole master's program is aimed at preparing the student to use research; the thesis is simply evidence of having achieved that goal.

A D 10. Many theses have had profound effect on subsequent research and products.

The question asked of you is this: "How do you *score* the responses given by an individual to this type of instrument?" Do you simply add the "agrees" to get a total score? What if some of the statements the subject agrees with are negative statements? Do you "reverse" the scoring for those items? How do you know which items are negative? Would a group of judges have the same opinion as yours as to which are positive or negative items?

Clearly, when measuring an attitude, there is no actual "correct" or "incorrect" response! In order to "score" an attitude instrument as that shown above, we must first establish the degree to which each item expresses an attitude that is favorable or unfavorable toward the "object" or topic for which the items are written. Some items when agreed with may give evidence of a very strong attitude toward the positive or the negative end of a continuum. If we can establish a *scale value* for each item that indicates the degree of "positiveness" toward the object, we can then use those scale values to score the responses of a subject. One of the ways of doing this is to use a group of "judges" to establish those scale values. The following illustrates an instrument used to garner the opinion of judges about the "positiveness" of the items in the previous instrument:

THESIS RESEARCH ATTITUDE INSTRUMENT
JUDGE EVALUATION FORM

DIRECTIONS:

You are being asked to determine the positiveness or negativeness each of the following items. To do this, you will rate each item on a scale ranging from 1 to 7 where 1 indicates highly negative to 7 which indicates highly positive. In order to have a common "frame of reference" for each item, assume that a graduate student has agreed with the statement, then rate how positive or negative that student is toward dissertation research. As an example, use the following item:

<u>A</u> D Most theses in Education are irrelevant surveys of little importance.

Assuming the student has marked AGREE (the underlined A) with the statement, how positive or negative do you think he (or she) is? Make a mark on the scale below to indicate your answer.

Highly Negative | Neither Positive Or Negative | Highly Positive

____1____|____2____|____3____|____4____|____5____|____6____|____7____

PLEASE BEGIN!

1. The research one does for his or her thesis may determine the line of research they pursue the rest of their life.

____1____|____2____|____3____|____4____|____5____|____6____|____7____

2. The only reason theses are required is because the current faculty had to do one in order to graduate.

____1____|____2____|____3____|____4____|____5____|____6____|____7____

3. Most theses make little contribution to the body of knowledge in a discipline.

____1____|____2____|____3____|____4____|____5____|____6____|____7____

4. A thesis can demonstrate your ability to be creative and thorough in

conducting a research project.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

5. Unless you almost have a major in statistics, its very difficult to complete a useful thesis.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

6. Reading a thesis is right up there with reading a telephone book for pleasure.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

7. Certain fields like clinical psychology, business and technology where the graduate is not going to be a college professor should not require a thesis.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

8. Ten years after completing their doctorate, most students are ashamed of their thesis.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

9. The whole doctorate program is aimed at preparing the student for research; the thesis is simply evidence of having achieved that goal.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

10. Many theses have had profound effect on subsequent research and products.
____1____|____2____|____3____|____4____|____5____|____6____|____7____

By analyzing the responses of a group of judges, the median or mean rating of those judges can be used to determine a scoring weight for each item that can be used in scoring the subjects for whom we wish to obtain an estimate of their attitude. One of the methods often used to analyze these judge's ratings is called the method of successive intervals (see Edwards, 1951). A computer program on you statistics disk permits you to analyze such responses. Consult the program manual for directions on its use.

Affective Measurement Theory

Most classroom teachers first learn to develop tests of achievement over the content which they are engaged to teach. These tests fall in what is known as the Cognitive Domain of testing. Two additional areas of testing are, however, often just as important. These areas are the Psychomotor Domain and the Affective Domain. The Psychomotor Domain includes testing of fine and gross motor coordination, strength and accuracy. The affective domain includes the measurement of attitudes, values and opinions of subjects. Typically, we are interested in measuring an attitude on one major "latent" variable such as an attitude toward school, an attitude toward minorities, an attitude toward some political issue, etc. In such cases, all of the items of the instrument used to measure this attitude are related, in some manner, to the major latent variable. In the following discussion, we will make this assumption of unidimensionality, that is, that all items are directly related to the same, underlying construct.

Thurstone Paired Comparison Scaling

A variety of item types have been developed to measure attitudes and values. Two major forms are used most commonly: (a) the agree/disagree format and (b) the "Likert" scale type involving a degree of agreement or disagreement, usually on a five or more point scale. In the case of agree/disagree statements, the subject is simply asked to indicate whether they agree or disagree to each statement listed. The statements are written to represent both positive or negative attitudes toward the object of the measurement. For example, if we were measuring an attitude toward "going to college" we might have the following statements:

1. College degrees are extremely important if your goal is to be a professional.
2. College graduates are snobish and have lost touch with humanity.
3. If you really want to make money, you can easily do so without going to college.
4. So many people are going to college, a college degree doesn't mean much any more.

If, on the other hand, we were using the Likert form of the statements, we will tell each subject to mark how strongly they agree (or disagree) with each statement using a scale such as

```
     _____|_____|_____|_____|_____|_____|_____|_____|_____
Strongly                                            Strongly
Disagree                                            Agree
```

You can see by the nature of the items, that there is no "correct" or "incorrect" response to each statement. Since we have no clear right or wrong answer, this poses a problem for "scoring" the responses of the instrument and obtaining a measure of the subject's attitude. We could arbitrarily mark those items which we feel reflect a positive attitude as a +1 if the subject "agreed" with the statement (or marked closer to the agree on a Likert scale), and score 0 if they failed to agree to a positive item. For negatively stated items we could similarly score a subject as 0 if they agreed with the negative item and score them a +1 if they disagreed with the negative item. The sum of these individual item scores, like our cognitive tests, would be the measure of the subject's attitude. Unfortunately, what you perceived as a "negative" or "positive" item may not be what I see for the same item! In fact, a group of judges might vary considerably in how "negative" or "positive" they felt each statement was toward the attitude object. Because of the ambiguity of attitude statements and because we desire to produce measurements for subjects which fit at least an interval scale of measurement, a variety of methods have been developed to "scale" the items used in affective instruments.

One of the first methods developed to determine the score values of items that subjects are asked to agree or disagree with is known as the Thurstone Paired-Comparisons Scaling method. This method utilizes a group of judges who are asked to compare each statement with every other statement and simply indicate which statement in each pair is more favorable toward the object if a subject were to agree with each one. For example, item 1 and item 2 of the above examples would be compared. If a judge felt that agreeing with item 1 indicated more favorableness toward going to college than agreeing with item 2, he would indicate item 1 is more favorable. By employing a reasonably large (say N > 20) number of judges, an average of the number of times judges selected each item over another can be obtained. If we assume these judgments by the judges are normally distributed around the "stimulus value" of each item, that is, the degree of favorableness of the items, we can obtain an estimate of the stimulus value for each item.

Let's consider an example of directions for the above 4 items that might be given to 30 judges:

```
DIRECTIONS:  Listed above are four statements which reflect
varying degrees of positiveness toward attending college.
Please indicate to the left of each pair of statements,
which item you feel reflects a more positive attitude toward
attending college.
_____ A.  Item 1     B.  Item 2
_____ A.  Item 1     B.  Item 3
_____ A.  Item 1     B.  Item 4
_____ A.  Item 2     B.  Item 3
_____ A.  Item 2     B.  Item 4
_____ A.  Item 3     B.  Item 4
```

Following administration of the above to 30 judges, we might obtain the following matrix. The number in the cells of this matrix reflect the number of judges which felt the item listed at the top was MORE favorable than the item listed to the left.

Judgement Matrix

ITEM	1	2	3	4
1	10	1	3	7
2	19	10	18	16
3	17	2	10	13
4	13	4	7	10

Notice in the above matrix that the diagonal values represent a comparison of a single item with itself. Since such comparisons are not actually made, we assume that one half of the time the item would be judged more positive and one half the time less positive. Also note that the values below the diagonal are the number of judges in the sample minus the value for the corresponding items above the diagonal.

To obtain the "scale value" of each item, we next convert the numbers of the above matrix first to the proportion of total judges and then we convert the proportions to z scores under the unit normal distribution. The matrices corresponding to the above example would be:

Proportion of Judgements

ITEM	1	2	3	4
1	.50	.05	.15	.35
2	.95	.50	.90	.80
3	.85	.10	.50	.15
4	.65	.20	.85	.50

z Scores for Proportions of Normal Curve

ITEM	1	2	3	4
1	0.00	-1.65	-1.04	-0.39
2	1.65	0.00	1.28	0.84
3	1.04	-1.28	0.00	-1.04
4	.39	- .84	1.04	0.00
Sum	3.08	-3.77	1.28	- .59
Average	.77	- .94	.32	- .15
Scale Value	1.71	0.00	1.26	.79

The last three rows above are simply the column sums, the column average, and the average plus the absolute value of the smallest column average. Since we are constructing a psychological scale, the mean and standard deviation of the scale values is arbitrary. We simply desire to build estimates of the intervals among the stimuli (items). The last row is labeled Scale Value. It reflects the average difference of the distance of each item from the other items on our psychological scale. The item (number 2) with the lowest scale value is the one which is "most negative" toward attending college. The item (number 1) with the largest value is the one most positive toward attending college. The scale values reflect the discriminations of the judges, NOT their attitudes. We simply used the judges to acquire "weights" for each item that reflect the degree of positivism or negativism of each item!

Now that we have these scale values however, we can use them to actually measure the attitude of subjects toward attending college. To do this, our subjects would receive instructions something like

```
Directions:  Each statement below reflects an attitude
               about college.  You are to circle the A if
               you agree with the statement or circle the D
               if you disagree with the statement.
               Go ahead.

     A    D    1.   College degrees are extremely important
                    if your goal is to be a professional.
     A    D    2.   College graduates are snobish and have
                    lost touch with humanity.
     A    D    3.   etc.
```

Once a subject has indicated agreement or disagreement with the items, the subject's total score is calculated by simply averaging the scale value of those items with which they agreed. The Paired-Comparisons procedure described above makes several assumptions. First, it assumes that the judges discriminations among the items are normally distributed. Secondly, it assumes that the variance of those discriminations are equal. Third, it assumes that the items all measure, to varying degrees, the same underlying attitude. Fourth, it assumes that the correlation among the judges discriminations for item pairs are all equal. Fifth, it assumes the mean and standard deviation of the scale values are arbitrary and the scale reflects only distances among items, not absolute amounts of an attitude.

You have probably already noticed that if you have very many items, the number of item pairs that judges are required to judge becomes large. The number of unique pairs is obtained by k(k-1)/2 where k is the number of items. For example, if you constructed 20 statements, the judges would have to make 20(19)/2 = 190 discriminations! Obviously you will try the patience of judges if your instrument is very long. A more convenient method of estimating item scale values is described in the next section.

Incidentally, if an item is judged to be higher than all other items by all judges or lower than all items by all judges, you would end up with a proportion of 1.0 or 0.0. The z scores corresponding to those proportions is plus or minus infinity and therefore could not be used to obtain an average. Such items may simply be eliminated or the obtained proportions changed to something like .99 or .01 as estimates of "what they might have been" if you had a much larger sample of judges.

Successive Interval Scaling Procedures

The Paired-Comparisons procedure described in the last section places great demands on judges if the number of items in an affective instrument is large. Yet we know that instruments with more items tend to give a more reliable estimate of an individual's attitude. The Successive Intervals scaling procedure provides a means of obtaining judges discriminations for k items in k judgments. The resulting scale values of items judged by both the Paired-Comparisons and Successive intervals methods correlate quite highly.

In the successive intervals scaling method, judges are asked to categorize statements on a continuum of an attribute like favorable-unfavorable. Typically five to nine categories are used, always using an odd number of categories. Utilizing the example from the previous section in which we are scaling items for measuring subjects attitudes toward attending college, a sample instruction to judges might look like the following:

```
Directions:  Each item below reflects some degree of
     favorableness or unfavorableness toward attending
     college.  Indicate the degree of favorableness in
     each item by making a check in one of the seven
     categories ranging from highly unfavorable to
     highly favorable.

1.   College degrees are extremely important if your
     goal is to be a professional.
     |_____|_____|_____|_____|_____|_____|_____|
     Highly                                Highly
     Unfavorable                           Favorable

2.   College graduates are snobbish and have lost
     touch with humanity.
     |_____|_____|_____|_____|_____|_____|_____|
     Highly                                Highly
     Unfavorable                           Favorable

3.   etc.
```

If we assume again that we have a reasonably large sample of judges evaluating each item of our instrument, and we assume that the classifications of items on the continuous scale tend to be normally distributed, we employ computations similar to the Paired-Comparison method for estimating scale values. For our example above, we might obtain, for the group of judges, the following classifications:

Frequency of Item Classifications

Category:	1	2	3	4	5	6	7
Item							
1	0	1	1	3	8	6	1
2	2	7	6	4	1	0	0
3	1	3	6	6	3	1	0
4	1	5	9	4	1	0	0

To obtain scale values by the method of successive intervals, we next obtain the cumulative frequencies within each item, convert those to cumulative proportions, and then convert the cumulative proportions to z scores. For example:

Cumulative Frequencies and Proportions

Category		1	2	3	4	5	6	7
Item:								
1	cf	0	1	2	5	13	19	20
	cp	0	.05	.10	.25	.65	.95	1.0
2	cf	2	9	15	19	20	20	20
	cp	.10	.45	.75	.95	1.0	1.0	1.0
3	cf	1	4	10	16	19	20	20
	cp	.05	.20	.50	.80	.95	1.0	1.0
4	cf	1	6	15	19	20	20	20
	cp	.05	.30	.75	.95	1.0	1.0	1.0

z Score Equivalents to Cumulative Proportions

Category	1	2	3	4	5	6	7
Item							

1	-	-1.65	-1.28	-0.67	0.38	1.65	-
2	-1.28	- .13	.68	1.65	-	-	-
3	-1.65	- .85	.00	.85	1.65	-	-
4	-1.65	- .52	.68	1.65	-	-	-

Differences Between Adjacent Categories

Difference	2-1	3-2	4-3	5-4	6-5	7-6
Item:						
1		.37	.61	1.05	1.27	
2	1.15	.81	.97			
3	.80	.85	.85	.80		
4	1.13	1.20	.97			
Sum	3.08	3.23	3.40	1.85	1.27	
N	3	4	4	2	1	
Mean	1.03	.81	.85	.93	1.27	
Cum. Avg.	1.03	1.84	2.69	3.62	4.89	

Scale values for the items which have been judged and analyzed by the method of successive intervals is obtained using the formula for the median of an interval, that is:

$$\text{Scale Value} = LL + \frac{(.5 - \Sigma Pb)}{Pw} \bar{W} \qquad (11.26)$$

where LL is the lower limit of the interval,
Pb is the Probability below the interval,
Pw is the Probability in the interval,

and $\bar{W}$ is the average interval width.

The scale value of items is the median value of the item on the scale defined by the cumulative average of the mean z score differences between categories. The scale values for the example above are therefore obtained as follows:

Scale value for item 1:

1. First, find the category in which the cumulative proportion is just less than .50, that is, that category just below the category in which the cumulative proportion is .5 or greater. For item 1 this is the category 4 (cumulative proportion = .25.
2. Next, obtain the cumulative average scale value for the category difference of the category just identified and the one below it. In this case, the cumulative average for the difference 4-3 which is 2.69. This represents the lower limit of the category in which the scale value for item 1 exists.
3. The value of ΣPb is the cumulative proportion up through the category identified in step (1) above, that is, .25 in our example.
4. The Pw is the proportion within the interval in which the median is found. In our example, it is the proportion obtained by subtracting the proportion up to category 5 from the proportion in category 5, that is, .65 - .25 = .40 .
5. Obtain the width of the interval next. This is the average z score differences in the interval in which the median is found. In this case the interval difference 5-4 has an average width of 0.93.
6. Substitute the values obtained in steps (1) - (5) in the equation to obtain the item scale value. For item 1 we have

$$S1 = 2.69 + \frac{(.50 - .25)}{(.65 - .25)} 0.93 = 3.2700$$

In a similar manner, the scale values for items 2 through 4 are:

$$S2 = 1.03 + \frac{(.50 - .45)}{(.75 - .45)} 0.81 = 1.1650$$

$$S3 = 1.03 + \frac{(.50 - .20)}{(.50 - .20)} 0.81 = 1.8343$$

$$S4 = 1.03 + \frac{(.50 - .30)}{(.75 - .30)} 0.81 = 1.3900$$

Several points should be made concerning the above computations. First note that the initial seven categories that were used represent midpoints of intervals. The number of judges placing an item within each category are assumed to be distributed uniformly accross the interval represented by the midpoint (category number). The calculation which involves subtracting the z scores in one category from those in the next higher category, and then averaging those values, establishes the distance between the midpoints of our original categories. In other words, there is no assumption of equal widths - we in fact estimate the interval widths. Once the interval widths are estimated, the accumulation of those widths describes the total scale of our measurements. You will have noticed that if the total number of categories is originally k (7 in our example), there will be k-2 differences obtained for adjacent categories. We have no way of estimating the width of the first and last category since there are no values below or above them. We can see this if we draw a schematic of the scale:

```
                          Midpoints
         1       2       3       4       5       6      7
_____|_______|_______|_______|_______|_______|_______|_____
         |_______|_______|_______|_______|_______|
             a       b       c       d       e
                          Intervals
```

We can illustrate where each item lies on the obtained scale by "plotting" the scale value of each item:

```
Item:          3   2 4              1
         |_______|_______|_______|_______|_______|
        0.0     1.0     2.0     3.0     4.0     5.0
```

We can see that item 1 was judged more positive than the other three items and lies considerably further from the other items. Items 2,3 and 4 are more similar in scale value with item 3 being judged the most negative of the four items.

Once the scale values of items are known, the same practice as employed in Paired-Comparisons methodology is used to obtain measures of individuals. The statements are presented to the subjects and the scale values of those items to which the subject agrees is averaged. The obtained average reflects the attitude of the subject.

Guttman Scalogram Analysis

If the items used to measure an attitude are all reflective of the same underlying attitude but to varying amounts, then subjects that vary on that attitude should agree or disagree to the items in a specific pattterm. As an example, assume we have 5 items which measure the degree of positivism toward maintaining U.S. troops in a base in Japan. Now assume that these items are ranked in the order to which they evoke an "agree" response by six people that vary in their attitude toward maintaining the troops in Japan. If there is consistency of measurement, and we assign a "1" if a subject "agrees" and "0" if the subject "disagrees" with an item, we would expect that the following matrix of observations might be recorded:

	Rank of Item on the Attitude						
	1	2	3	4	5	Score	Rank
Subject							
1	1	1	1	1	1	5	1
2	0	1	1	1	1	4	2
3	0	0	1	1	1	3	3
4	0	0	0	1	1	2	4
5	0	0	0	0	1	1	5
6	0	0	0	0	0	0	6

In our example, subject 1 has agreed with all five statements and subject 6 has disagreed with all items. Note the items have been arranged in order from most negative toward maintaining troops to most positive toward retaining troops in Japan. In addition, the subjects have been arranged from the subject with the most positive attitude down to the subject with the least positive (most negative) attitude. The matrix of the responses reflects perfect agreement or order of the responses. In "real" life, we seldom get such a perfect pattern of responses. A more typical response pattern might look more like:

	Items Ordered by Total "Agree" Responses										
	1		2		3		4		5		
Response	1	0	1	0	1	0	1	0	1	0	Score
Subject											
1	x		x		x		x		x		5
2		x	x		x		x		x		4
3	x			x	x		x		x		4
4		x	x			x	x		x		3
5		x		x		x	x		x		2
6		x		x	x			x	x		2
7		x		x		x		x	x		1
8		x		x		x	x				1
9		x		x		x		x		x	0
10		x		x		x		x		x	0
sums	2	8	3	7	4	6	6	4	7	3	
Proportion	.2	.8	.3	.7	.4	.6	.6	.4	.7	.3	

In this sample of ten subjects, we have several subjects with the same total score as another subject but a different pattern of "agree" or "disagree" to the statements. There is not perfect agreement among the items in differentiating the attitudes of the subjects! Note that we have recorded the response of each subject in one of two columns beneath each item. The sum or proportion of the "agree" or 1 responses is totaled accross subjects to identify the order of the "positivism" of the item. Item 5 is the item which received the greatest number of "agree" responses while item 1 received the fewest.

If we have "perfect" reproducibility in an instrument of k items, we would be able to perfectly reproduce the individual item responses of an individual given their total score (number of items to which they agree). If their is inconsistency of measurement, we can only estimate the likely response to each item. In order to make such estimates, it is necessary to identify a "cutting" point for each item which identifies that point where the pattern of agree/disagree responses most likely changes. This point is one where the number of errors is a minimum. An error is counted whenever a subject below the cutting score agrees with a statement or whenever a subject above the cutting point disagrees with the statement. For the above table, we have inserted the cutting scores which give the minimum error counts:

	Items Ordered by Total "Agree" Responses				
	1	2	3	4	5

```
Response     1 0     1 0     1 0     1 0     1 0     Score

Subject
   1         x       x       x       x       x         5
   2           x     x       x       x       x         4
   3         x__       x     x__     x       x         4
   4           x     x__       x     x       x         3
   5           x       x       x     x__     x         2
   6           x       x     x         x     x         2
   7           x       x       x       x     x__       1
   8           x       x       x     x         x       1
   9           x       x       x       x       x       0
  10           x       x       x       x       x       0

sums         2 8     3 7     4 6     6 4     7 3
Proportion .2  .8  .3  .7  .4  .6  .6  .4  .7  .3
Errors       0 1     0 1     1  0    1 0     0 0    Σe=4
```

There are actually several choices for cutting scores on each item which minimize the sum of the errors. L. Guttman (see Edwards, p. 182) has developed a coefficient which expresses the degree of reproducibility of a set of items. It is obtained as one minus the proportion of errors in the total number of responses. For the above data, we would obtain the coefficient of reproducibility as

$$Rep = 1.0 - 4/50 = 0.92$$

Because the cutting scores in the above matrix may be made at several points, the response pattern expected of a subject with a given total score might vary from solution to solution. In order to obtain a method of setting cutting scores that is always the same and thus yields a means of accurately predicting a response pattern, Edwards (Edwards, pgs. 184-188) developed another method for obtaining cutting scores. This method is illustrated for the same data in the Fig. below:

```
          Items Ordered by Total "Agree" Responses
              1       2       3       4       5
Response     1 0     1 0     1 0     1 0     1 0     Score

Subject
   1         x       x       x       x       x         5
   2           x     x       x       x       x         4
           -------------------------------------
   3         x         x     x       x       x         4
           -------------------------------------
   4           x     x         x     x       x         3
           -------------------------------------
   5           x       x       x     x       x         2
   6           x       x     x         x     x         2
           -------------------------------------
   7           x       x       x       x     x         1
           -------------------------------------
   8           x       x       x     x         x       1
   9           x       x       x       x       x       0
  10           x       x       x       x       x       0

sums         2 8     3 7     4 6     6 4     7 3
Proportion .2  .8  .3  .7  .4  .6  .6  .4  .7  .3
```

In the above display of our sample data, we have used the proportion of 1 responses (agree) to draw our cutting points. For example, in item 1, 20 percent of the subjects agreed with the item. The cutting score was then drawn below 20 percent of all the responses (both agree and disagree). This procedure was used for each item.

Errors are then counted whenever a response disagrees with the pattern expected. For example, both subjects 1 and 2 are expected to have a pattern of responses 1 1 1 1 1 but subject 2 has 0 1 1 1 1 as a pattern. One response disagreed with the expected so the error count is 1 for subject 2. Subject three is expected to have a response pattern of 0 1 1 1 1 but in fact has a response pattern of 1 0 1 1 1 . Since there are two items that disagree with the expect pattern, the error count for subject 3 is 2. A similar procedure is followed for each subject. The expected pattern for each total score is shown below along with the number of errors counted for subjects with those total scores:

Total Score	Expected Pattern	Subject	No. of Errors
5	1 1 1 1 1	1	0
4	0 1 1 1 1	2	0
		3	2
3	0 0 1 1 1	4	2
2	0 0 0 1 1	5	0
		6	2
1	0 0 0 0 1	7	0
		8	2
0	0 0 0 0 0	9	0
		10	0
			$\Sigma e = 8$

$$Rep = 1.0 - (8 / 50) = 0.84$$

This computation of the coefficient of reproducibility is a measure of the degree of accuracy with which statement responses can be reproduced on the basis of the total score alone! It is this latter method with is used in the program GUTTMAN found in the LazStats program. The proportion of subjects agreeing or disagreeing with each item affects the degree of reproducibility. If very large or very small numbers of subjects agree to an item, the reproducibility is increased. The minimal coefficient of reproducibility may be obtained by the larger of the two values (a) proportion agreeing or (b) proportion disagreeing with a statement and dividing by the number of items. In our example these values are .8, .7, .6, .6 and .7. The minimal marginal reproducibility is therefore

$$\frac{.8 + .7 + .6 + .6 + .7}{6} = 0.68$$

The response pattern corresponding to this model response pattern is 0 0 0 1 1 . If we were to predict each subjects responses with this pattern and count errors, the coefficient of reproducibility would be .68! The Guttman Coefficient of reproducibility may be thought of as an index somewhat comparable to the reliability coefficient. A value of one would indicate a set of items that are fully consistent in measuring differences among subjects.

In the methods of paired comparison and successive intervals, we utilized a group of judges to estimate scale values for items. These scale values were then used to obtain the scores for subjects administered the statements. With the Guttman scaling method, we do not use judges but simply the responses of the subjects themselves as a basis for determining their attitude scores. We simply assign 1 to the item with which they agree and 0 to those with which they disagree. If the instrument has a high coefficient of reproducibility, then the total of the subject response codes, i.e. their total score, should be directly interpretable as a measure of their attitude. The subject's total score may be divided by the number of items to obtain the proportion of items to which the subject agreed. It is assumed that all items reflect a varying degree of positivism to the attitude object (e.g. troops in Japan) and therefore the subject's total score based on those items also reflects the subject's attitude. The scale value of each item is the cutting score for that item. In the above example, we may place the items on the scale as follows:

```
Item          1   2   3       4   7
     |_______|_______|_______|_______|_______|
     0      .2      .4      .6      .8      1.0
         Proportion of "Agree" Items
```

The items to which few subjects "agree" is a more negative item than the item to which a larger number of subjects agree. The proportion of items an individual subject agrees with is an indication of the subjects positivism toward the attitude object.

Likert Scaling

Also called the method of Summated Ratings, the Likert scaling method, like the Guttman method above, does not use judges to determine the scale value of items. Subjects are directly measured on each statement by indicating their degree of agreement, usually using a five-point scale. The statements administered are statements judged only by the person constructing the items as either a "favorable" or "unfavorable" item. If a five point scale is used such as

```
|________|________|________|________|________|
Strongly                             Strongly
Disagree                             Agree
```

the lowest category is assigned a value of 0, the next category a 1, etc. up to the last category which would be assigned the value 4. If the item is an "unfavorable" item toward the attitude object, the category scores are reversed, that is, the first category assigned 4, the next 3, etc. To obtain a subject's score, one simply adds the values of the categories checked by the subject. Normal item analysis procedures may be used to eliminate items which do not measure the attitude consistent with other items. The point-biserial correlation of the item with the total score is the typical criterion used. If the item correlates quite low with the total score, the item should be eliminated.

It is important to note that the scores obtained by the Likert method cannot be interpreted without reference to a comparison group. Since the item scale values are not obtained, and the distances among the items is therefore unknown, the total scores are only meaningful in reference to a comparison group. For example, say that a scale of 20 items is administered to a subject and the subject's score is 5. This score cannot be directly interpreted. It may be that in one group of subjects this is a highly positive score while in another group, a very low score. We cannot say the score of 5, by itself, reflects a positive or negative attitude toward the object. It has been found in previous research that scores obtained on a Likert scale correlate quite high with the same items scaled and scored by the Thurstone method. If the interest of the researcher is to use the attitude measures to describe its relationship with some other variables through correlation methods, then the Likert method is cost-effective. If, on the other hand, the researcher desires to interpret individual attitudes as being positive or negative toward some object, then a method such as the paired-comparison or successive interval scaling method should be employed.

Semantic Differential Scales

Osgood, et al (1971) developed a measure of the "meaning" attached, through a theorized learning model, to a variety of stimuli including both physical objects as well as "ideas" or concepts. Their measure is based, briefly, on the notion that certain words have become associated with subject's responses to objects through conditioning and generalization of conditioning. They observed that in many situatations, people, for example, might use words such as heavy, dark, gloomy to describe some classical music while words such as bright, up, shiny, happy might describe other music. These words which are also used to describe many objects appear to have general utility for subjects in describing their "feelings" about an object. Osgood and his colleagues utilized factor analysis procedures to identify subsets of items which appear to measure different dimensions of meaning. Their goal was to identify a set of bipolar adjectives which describes the "semantic space" of given objects. This space is described by orthogonal axis of the bipolar adjectives. The objects lie within this space at varying distance from the origin (intensity) and in specific directions (description). Three major dimensions of the semantic space are typically used. These are (I) Evaluation, (II) Activity, and (III) Potency.

The semantic differential scale is constructed of those bipolar adjectives (e.g. hot - cold) which are demonstrated to differentiate the meaning attached by individuals to a given object (e.g. school attendance). Thus the first problem in constructing a semantic differential scale is the selection of bipolar adjective pairs that measure predominantly one dimension of the semantic space and differentiate among individuals that vary in intensity of feeling on that dimension. Once the adjectives have been identified and their discriminating potential demonstrated, the selected items are utilized to measure the feelings (attitudes or values) that individual subjects attach to the object.

Typical instructions to subjects are as follows:

Directions: This instrument is designed to measure the meaning of certain things by having people judge them with a series of scales using word opposites. Make your judgments on the basis of what these things mean to YOU. Below you will see the thing to be judged in the center of the page. You are to rate this object on each of the scales below the object. Here is how you use the scales:

If you feel the object in the center is very closely related to one end of the scale, you should place your check-mark as follows:

fair__X__|_____|_____|_____|_____|_____|_____unfair

or

fair_____|_____|_____|_____|_____|_____|__X__unfair

If you feel the concept is quite closely related to one or the other end of the scale (but not extremely), you should place your check-mark as follows:

strong_____|__X__|_____|_____|_____|_____|_____weak

or

strong_____|_____|_____|_____|_____|__X__|_____weak

If the object seems only slightly related to one side as opposed to the other side (but is really not neutral), then you should check as follows:

active_____|_____|__X__|_____|_____|_____|_____passive

or

active_____|_____|_____|_____|__X__|_____|_____passive

If you consider the concept to be neutral on the scale, both sides equally associated with the object, or if the scale is completely irrelevant, unrelated to the concept, then you should place your check-mark in the middle space:

safe_____|_____|_____|__X__|_____|_____|_____dangerous

GO AHEAD!

SCHOOL

1. good _____|_____|_____|_____|_____|_____|_____ bad

2. kind _____|_____|_____|_____|_____|_____|_____ cruel

3. high _____|_____|_____|_____|_____|_____|_____ low

```
4.   hard _____|_____|_____|_____|_____|_____|_____ soft

5.   heavy_____|_____|_____|_____|_____|_____|_____ light

6.   sane _____|_____|_____|_____|_____|_____|_____ insane

7.   near _____|_____|_____|_____|_____|_____|_____ far

8.   etc.
```

Typically, 3 or more items are selected from those items which "load" heaviest on each of the factors or dimensions of the semantic space which the researcher wishes to measure. More items from a given dimension yields a more reliable estimate of that dimension. Note that if items from more than one factor are used, a profile of scores may be obtained for each individual. The user of the semantic differential scales may choose, of course, to measure on only one dimension. Items may also be included that are not previously known to load on a particular dimension but are felt by the test constructor to be relevant for measureing the meaning or attitude toward a given object. Later analyses may then be performed to determine the extent to which these other items load on the dimensions of the semantic space.

While it is assumed that the scales (items) of the semantic differential scales are equal interval scales, this assumption may be checked by using the successive interval scaling program to estimate the interval widths of the individual items. Dimension scores for individuals are usually computed by simply summing or averaging the scale values of each item where the scale values are -3, -2, -1, 0, +1, +2 and +3 corresponding to the seven categories used. Notice that the values may need to be reversed if the "negative" synonym is listed first and the "positive" listed last.

Behavior Checklists

The industrial technology evaluator will sometimes utilize a behavior checklist form to record observations regarding work habits, verbal interactions, or events considered important to a given study. In industrial training situations, the evaluator may record such details as the number of steps taken during a given operation, the frequency of lifting objects from below waist level, the number of manual adjustments to equipment, etc. related to the training. Time and motion studies may provide valuable information for reducing fatigue and injury, reducing operating times for processes, and suggest alternative methods of operation. In evaluating trainer performance, a behavior checklist may “zero in” on specific behaviors potentially detracting from the effectiveness of the instructor as well as identifying those important to retain and reinforce.

As an example of a behavioral checklist, consider the following set of “items” by which trainees record their observations about behaviors of a trainer:

Behavior of the Trainer

Directions: Each item below describes a behavior that you might have observed during the training session. For each item indicate whether or not the behavior occurred and indicate how you felt about the behavior. Express your feeling about the behavior by checking one of the numbers between 1 and 5 where 1 indicates “Highly undesirable”, 2 indicates slightly undesirable, 3 indicates neither desirable or undesirable, 4 indicates somewhat desirable and 5 indicates “Definitely desirable”.

	ITEM	**OBSERVED? (Y OR N)**	**FEELING 1**	**2**	**3**	**4**	**5**
1.	Embarrassed a trainee.	_____________	__	__	__	__	__
2.	Arrived late for a session.	_____________	__	__	__	__	__
3.	Showed enthusiasm for the subject.	_____________	__	__	__	__	__
4.	Showed a good sense of humor.	_____________	__	__	__	__	__
5.	Showed sensitivity to the learner.	_____________	__	__	__	__	__

6. Got off the subject. ____________ __ __ __ __ __
7. Talked over my head. ____________ __ __ __ __ __
8. Reviewed what we had learned. ____________ __ __ __ __ __
9. Handed out helpful reading material. ____________ __ __ __ __ __
10. Used inappropriate English. ____________ __ __ __ __ __

To "score" the above type of data, the evaluator may multiply the value of the "feeling" scale checked by one (1) if the observer marked "y" to observing it or zero (0) if not observed. The higher the score, the "better" the trainer behaved in the view of the trainees.

Codifying Personal Interactions

In some situations, it is necessary to evaluate the content of interpersonal communications. For example, to create a work environment free of discrimination, the conversations among employees may be coded for words, phrases, sentences, gestures, or behaviors which may be construed as sexist, discriminatory or derogatory to other individuals. Unfortunately, one cannot always sit and take notes while others are conversing. Use of tape recording without the permission of those recorded is also inappropriate. Often the best one can do is to take note of a part of a conversation overheard, record one's observations as soon as possible afterwards, and then, if possible, verify what was heard with one or more persons that may also have heard the conversation. Clearly, this is an emotionally laden and sensitive area! One must use extremely good judgment. Rather than recording specific "offenders" names, for example, one may use code letters or numbers to represent individuals. One may also encode words, gestures, etc. within categories. Let's consider an example where a female employee has complained of sexual harassment in a business which employs primarily men and very few women in packaging meat for retail store distribution. A consultant is hired to evaluate the work place for evidence of a problem with sexual harassment. The evaluator first does a "walk-through" to garner any graphical evidence of harassment such as :

g1 = sexually explicit graffiti or pictures in view in restrooms
g2 = written material making explicit sexual innuendoes regarding an employee

Next, the evaluator may draw a random sample of employees and formally interview them, giving full assurance of confidentiality. The evaluator may code each employees responses as E1, E2, etc. and, using a pre-defined schedule of questions, code the responses to each question as + or - to indicate statements made that verify or negate the presence of harassment. Again, the coding for the questions and their responses might be:

E1(1) +; E1(2) -; E1(3)-; E1(4)+
E2(1)-; E2(2)+; E2(3)+; E2(4)-
etc.

The evaluator may specifically interview the females in the work-setting (recognizing that sexual harassment *can* be evidenced by either gender, but more likely reported by females). This type of interview is again, very sensitive. An individual often must show great courage to even raise the complaint of harassment and may fear reprisal from coworkers or employer. The evaluator must be particularly well versed in the separation of *perceptions* of harassment from *evidence* of harassment. Again, coding of responses to questions or volunteered information may be useful for assuring confidentiality and brevity in data collection. Something like the previous coding might be used:

C1(1)V+; C1(2)P-; etc. where C1 is the first complainant, V is evidence, P is a perception and + or - is content within the definition of harassment or not in the definition of harassment.

Once such data is collected and summarized, the evaluator must still attach weight to each type of evidence or perception. Typically, "hard" evidence such as graffiti, derogatory written comments, verified derogatory conversations, etc. are given a higher value than perceptions or hearsay evidence. Notice that the evaluator is not in the role of changing the work environment, filing complaints with the Equal Opportunity Commission or other corrective decisions and actions. The evaluator in this example was likely asked to determine if harassment exists or perhaps the "degree" of harassment that may exist. The report completed may, of course, suggest alternative actions appropriate to the evidence found and conclusions reached by the evaluator. It is the responsibility of the evaluators employer to act on the evaluation results, not the evaluator.

Classical Test Item Analysis

Classical item analysis is used to estimate the reliability of test scores obtained from measures of subjects on some attribute such as achievement, aptitude or intelligence. In classical test theory, the obtained score for an individual on items is theorized to consist of a "true score" component and an "error score" component. Errors are typically assumed to be normally distributed with a mean of zero over all the subjects measured.

Several methods are available to estimate the reliability of the measures and vary according to the assumptions made about the scores. The Kuder-Richardson estimates are based on the product-moment correlation (or covariance) among items of the observed test scores and those of a theoretical "parallel" test form. The Cronbach and Hoyt estimates utilize a treatment by subjects analysis of variance design which yields identical results to the KR#20 method when item scores are dichotomous (0 and 1) values.

When you select the Classical Item Analysis procedure you will use the following dialogue box to specify how your test is to be analyzed. If the test consists of multiple sub-tests, you may define a scale for each sub-test by specifying those items belonging to each sub-test. The procedure will need to know how to determine the correct and incorrect responses. If your data are already 0 and 1 scores, the most simple method is to simply include, as the first record in your file, a case with 1's for each item. If your data consists of values ranging, say, between 1 and 5 corresponding to alternative choices, you will either include a first case with the correct choice values or indicate you wish to Prompt for Correct Responses (as numbers when values are numbers.) If items are to be assigned different weights, you can assign those weights by selecting the "Assign Item Weights scoring option. The scored item matrix will be printed if you elect it on the output options. Three different reliability methods are available. You can select them all if you like.

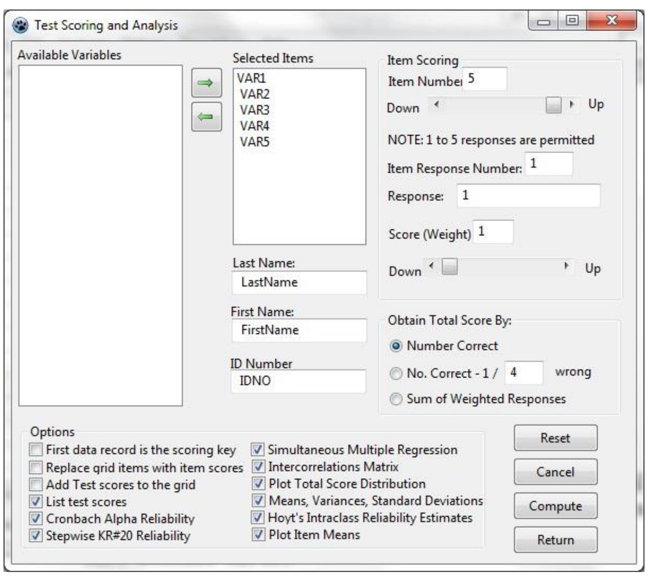


Fig. 11.1   Classical Item Analysis Dialog

Shown below is a sample output obtained from the Classical Item Analysis procedure followed by an item characteristic curve plot for one of the items. The file used was "itemdat.LAZ".

```
TEST SCORING REPORT

PERSON ID NUMBER FIRST NAME LAST NAME TEST SCORE
               1       Bill     Miller   5.00
               2       Barb     Benton   4.00
               3        Tom   Richards   3.00
               4      Keith     Thomas   2.00
```

```
               5        Bob       King   1.00
               6        Rob   Moreland   0.00
               7      Sandy     Landis   1.00
               8     Vernil      Moore   2.00
               9       Dick      Tyler   3.00
              10      Harry       Cook   4.00
              11     Claude      Rains   5.00
              12      Clark       Kent   3.00
              13       Bill    Clinton   3.00
              14     George       Bush   4.00
              15        Tom  Jefferson   4.00
              16        Abe    Lincoln   2.00
Alpha Reliability Estimate for Test = 0.6004  S.E. of Measurement =    0.920
Analysis of Variance for Hoyt Reliabilities

SOURCE    D.F.          SS        MS        F        PROB
Subjects   15          6.35      0.42      2.50      0.01
Within     64         13.20      0.21
Items       4          3.05      0.76      4.51      0.00
Error      60         10.15      0.17
Total      79         19.55

Hoyt Unadjusted Test Rel. for scale TOTAL  = 0.5128  S.E. of Measurement =    0.000
Hoyt Adjusted Test Rel. for scale TOTAL    = 0.6004  S.E. of Measurement =    0.000
Hoyt Unadjusted Item Rel. for scale TOTAL  = 0.1739  S.E. of Measurement =    0.000
Hoyt Adjusted Item Rel. for scale TOTAL    = 0.2311  S.E. of Measurement =    0.000
Item and Total Score Intercorrelations with   16 cases.


Variables
                VAR1        VAR2        VAR3        VAR4        VAR5
     VAR1       1.000       0.153       0.048      -0.048       0.255
     VAR2       0.153       1.000       0.493       0.323       0.164
     VAR3       0.048       0.493       1.000       0.270       0.323
     VAR4      -0.048       0.323       0.270       1.000       0.221
     VAR5       0.255       0.164       0.323       0.221       1.000
     TOTAL      0.369       0.706       0.727       0.615       0.634


Variables
               TOTAL
     VAR1       0.369
     VAR2       0.706
     VAR3       0.727
     VAR4       0.615
     VAR5       0.634
     TOTAL      1.000




Means with   16 valid cases.

Variables       VAR1        VAR2        VAR3        VAR4        VAR5
                0.875       0.688       0.563       0.438       0.313

Variables       TOTAL
                2.875


Variances with   16 valid cases.

Variables       VAR1        VAR2        VAR3        VAR4        VAR5
                0.117       0.229       0.263       0.263       0.229

Variables       TOTAL
                2.117


Standard Deviations with   16 valid cases.

Variables       VAR1        VAR2        VAR3        VAR4        VAR5
                0.342       0.479       0.512       0.512       0.479
```

```
Variables        TOTAL
                 1.455
KR#20 = 0.6591 for the test with mean =  1.250 and variance =  0.733
Item  Mean    Variance   Pt.Bis.r
  2    0.688   0.229     0.8538
  3    0.563   0.263     0.8737
KR#20 = 0.6270 for the test with mean =  1.688 and variance =  1.296
Item  Mean    Variance   Pt.Bis.r
  2    0.688   0.229     0.7875
  3    0.563   0.263     0.7787
  4    0.438   0.263     0.7073
KR#20 = 0.6310 for the test with mean =  2.000 and variance =  1.867
Item  Mean    Variance   Pt.Bis.r
  2    0.688   0.229     0.7135
  3    0.563   0.263     0.7619
  4    0.438   0.263     0.6667
  5    0.313   0.229     0.6116
KR#20 = 0.6004 for the test with mean =  2.875 and variance =  2.117
Item  Mean    Variance   Pt.Bis.r
  2    0.688   0.229     0.7059
  3    0.563   0.263     0.7267
  4    0.438   0.263     0.6149
  5    0.313   0.229     0.6342
  1    0.875   0.117     0.3689
Item and Total Score Intercorrelations with   16 cases.


Variables
                  VAR1         VAR2         VAR3         VAR4         VAR5
     VAR1        1.000        0.153        0.048       -0.048        0.255
     VAR2        0.153        1.000        0.493        0.323        0.164
     VAR3        0.048        0.493        1.000        0.270        0.323
     VAR4       -0.048        0.323        0.270        1.000        0.221
     VAR5        0.255        0.164        0.323        0.221        1.000
     TOTAL       0.369        0.706        0.727        0.615        0.634


Variables
                  TOTAL
     VAR1        0.369
     VAR2        0.706
     VAR3        0.727
     VAR4        0.615
     VAR5        0.634
     TOTAL       1.000



Means with   16 valid cases.

Variables        VAR1         VAR2         VAR3         VAR4         VAR5
                 0.875        0.688        0.563        0.438        0.313

Variables        TOTAL
                 2.875


Variances with   16 valid cases.

Variables        VAR1         VAR2         VAR3         VAR4         VAR5
                 0.117        0.229        0.263        0.263        0.229

Variables        TOTAL
                 2.117


Standard Deviations with   16 valid cases.

Variables        VAR1         VAR2         VAR3         VAR4         VAR5
                 0.342        0.479        0.512        0.512        0.479

Variables        TOTAL
                 1.455
```

Determinant of correlation matrix = 0.5209

Multiple Correlation Coefficients for Each Variable

Variable	R	R2	F	Prob.>F	DF1	DF2
VAR1	0.327	0.107	0.330	0.852	4	11
VAR2	0.553	0.306	1.212	0.360	4	11
VAR3	0.561	0.315	1.262	0.342	4	11
VAR4	0.398	0.158	0.516	0.726	4	11
VAR5	0.436	0.190	0.646	0.641	4	11

Betas in Columns with 16 cases.

Variables

	VAR1	VAR2	VAR3	VAR4	VAR5
VAR1	-1.000	0.161	-0.082	-0.141	0.262
VAR2	0.207	-1.000	0.442	0.274	-0.083
VAR3	-0.107	0.447	-1.000	0.082	0.303
VAR4	-0.149	0.226	0.067	-1.000	0.178
VAR5	0.289	-0.071	0.257	0.185	-1.000

Standard Errors of Prediction

Variable	Std.Error
VAR1	0.377
VAR2	0.466
VAR3	0.495
VAR4	0.549
VAR5	0.503

Raw Regression Coefficients with 16 cases.

Variables

	VAR1	VAR2	VAR3	VAR4	VAR5
VAR1	-1.000	0.225	-0.123	-0.211	0.367
VAR2	0.147	-1.000	0.473	0.293	-0.083
VAR3	-0.071	0.418	-1.000	0.082	0.283
VAR4	-0.099	0.211	0.067	-1.000	0.167
VAR5	0.206	-0.071	0.275	0.199	-1.000

Variable	Constant
VAR1	0.793
VAR2	0.186
VAR3	0.230
VAR4	0.313
VAR5	-0.183

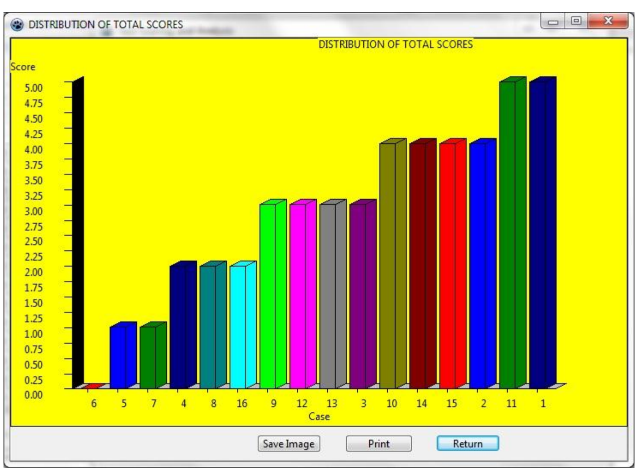


Fig. 11.2 Distribution of Test Scores (Classical Analysis)

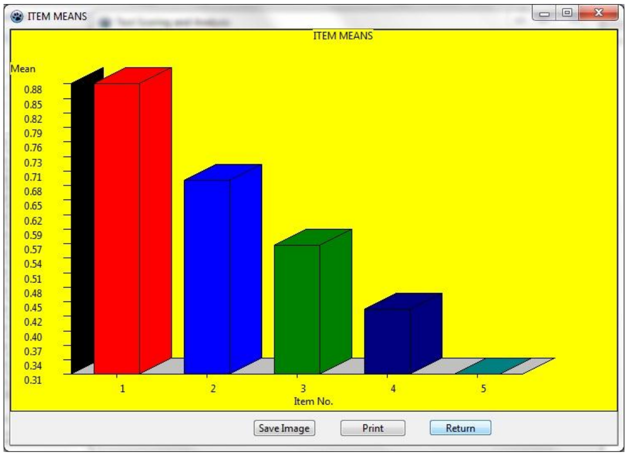


Fig. 11.3 Item Means

Analysis of Variance: Treatment by Subject and Hoyt Reliability

The Within Subjects Analysis of Variance involves the repeated measurement of the same unit of observation. These repeated observations are arranged as variables (columns) in the Main Form grid for the cases (grid rows.) If only two measures are administered, you will probably use the matched pairs (dependent) t-test method. When more than two measures are administered, you may use the repeated measures ANOVA method to test the equality of treatment level means in the population sampled. Since within subjects analysis is a part of the Hoyt Intraclass reliability estimation procedure, you may use this procedure to complete the analysis (see the Measurement procedures under the Analyses menu on the Main Form.)

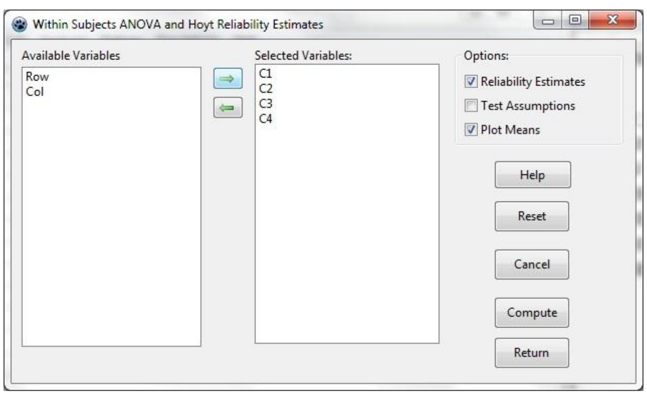


Fig. 11.4 Hoyt Reliability by ANOVA

```
The output from an example analysis is shown below:

Treatments by Subjects (AxS) ANOVA Results.

Data File = C:\lazarus\Projects\LazStats\LazStatsData\ABRDATA.LAZ


-----------------------------------------------------------
SOURCE           DF        SS        MS        F  Prob. > F
-----------------------------------------------------------
SUBJECTS         11   181.000    16.455
WITHIN SUBJECTS  36  1077.000    29.917
   TREATMENTS     3   991.500   330.500   127.561     0.000
   RESIDUAL      33    85.500     2.591
-----------------------------------------------------------
TOTAL            47  1258.000    26.766
-----------------------------------------------------------
```

```
TREATMENT (COLUMN) MEANS AND STANDARD DEVIATIONS
VARIABLE  MEAN      STD.DEV.
C1          16.500     2.067
C2          11.500     2.431
C3           7.750     2.417
C4           4.250     2.864

Mean of all scores =     10.000 with standard deviation =      5.174

RELIABILITY ESTIMATES

TYPE OF ESTIMATE              VALUE
Unadjusted total reliability  -0.818
Unadjusted item reliability    -0.127
Adjusted total (Cronbach)      0.843
Adjusted item reliability      0.572

BOX TEST FOR HOMOGENEITY OF VARIANCE-COVARIANCE MATRIX


SAMPLE COVARIANCE MATRIX with   12 cases.


Variables
                     C1            C2            C3            C4
        C1       4.273         2.455         1.227         1.318
        C2       2.455         5.909         4.773         5.591
        C3       1.227         4.773         5.841         5.432
        C4       1.318         5.591         5.432         8.205




ASSUMED POP. COVARIANCE MATRIX with   12 cases.


Variables
                     C1            C2            C3            C4
        C1       6.057         0.693         0.693         0.693
        C2       0.114         5.977         0.614         0.614
        C3       0.114         0.103         5.914         0.551
        C4       0.114         0.103         0.093         5.863


Determinant of variance-covariance matrix =       81.6
Determinant of homogeneity matrix =   1.26E003
ChiSquare =    108.149 with   8 degrees of freedom
Probability of larger chisquare = 9.66E-007
```

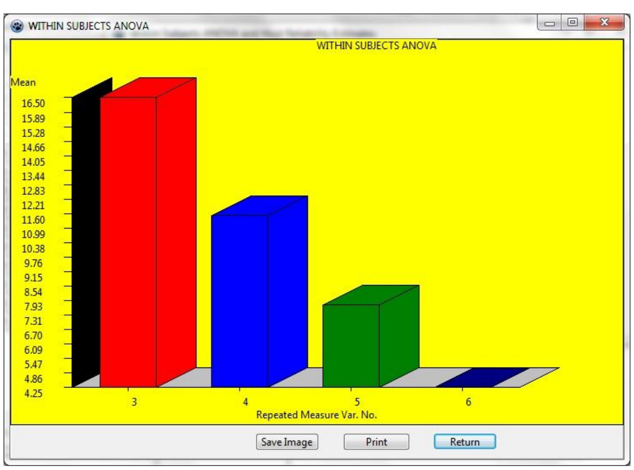


Fig. 11.5   Within Subjects ANOVA Plot

Kuder-Richardson #21 Reliability

The Kuder-Richardson formula #20 was developed from Classical Test Theory (true-score theory). A shorter form of the estimate can be made using only the mean, standard deviation and number of test items if one can assume that the inter-item covariances are equal. Below is the form which appears when this procedure is selected from the Measurement option of the Analyses menu:

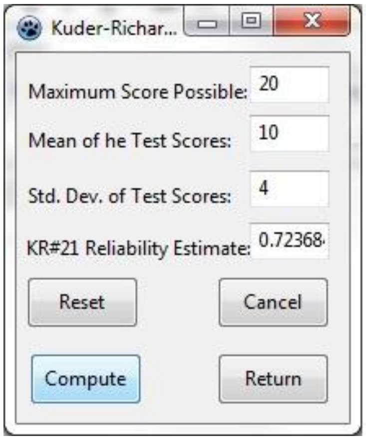


Fig. 11.6 Kuder-Richardson Formula 21 Reliability

Note that we have entered the maximum score (total number of items), the test mean, and the test standard deviation. When you click the Compute button, the estimate is shown in the labeled box.

Weighted Composite Test Reliablity

The reliability for a combination of tests, each of which has its own estimate of reliability and a weight assigned to it, may be computed. This composite will typically be greater than any one test by itself due to the likelihood that the subtests are correlated positively among themselves. Since teachers typically assign course grades based on a combination of individual tests administered over the time period of a course, this reliability estimate in built into the Grading System. See the description and examples in that section. A file labeled "CompRel.LAZ" is used in the example below:

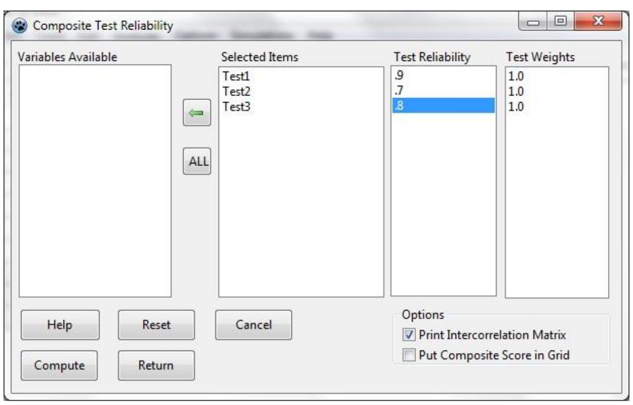


Fig. 11.7 Composite Test Reliability Dialog

```
Composite Test Reliability

File Analyzed: C:\lazarus\Projects\LazStats\LazStatsData\CompRel.LAZ
```

```
Correlations Among Tests with   10 cases.


Variables
                  Test1        Test2        Test3
     Test1       1.000        0.927        0.952
     Test2       0.927        1.000        0.855
     Test3       0.952        0.855        1.000




Means with   10 valid cases.

Variables        Test1        Test2        Test3
                 5.500        5.500        7.500


Variances with   10 valid cases.

Variables        Test1        Test2        Test3
                 9.167        9.167        9.167


Standard Deviations with   10 valid cases.

Variables        Test1        Test2        Test3
                 3.028        3.028        3.028



Test Weights with   10 valid cases.

Variables        Test1        Test2        Test3
                 1.000        1.000        2.000


Test Reliabilities with   10 valid cases.

Variables        Test1        Test2        Test3
                 0.900        0.700        0.800

Composite reliability =  0.929
```

Rasch One Parameter Item Analysis

Item Response Theory (IRT) is another theoretical view of subject responses to items on a test. IRT suggests that items may posess one or more characteristics (parameters) that may be estimated. In the theory developed by George Rasch, one parameter, item difficulty, is estimated (in addition to the estimate of individual subject “ability” parameters.) Utilizing maximum-liklihood methods and log difficulty and log ability parameter estimates, the Rasch method attempts to estimate subject and item parameters that are “independent” of one another. This is unlike Classical theory in which the item difficulty (proportion of subects passing an item) is directly a function of the ability of the subjects sampled. IRT is sometimes also considered to be a “Latent Trait Theory” due to the assumption that all of the items are measures of the same underlying “trait”. Several tests of the “fit” of the item responses to this assumption are typically included in programs to estimate Rasch parameters. Other IRT procedures posit two or three parameters, the others being the “slope” and the “chance” parameters. The slope is the rate at which the probability of getting an item correct increases with equal units of increase in subject ability. The chance parameter is the probability of obtaining the item correct by chance alone. In the Rasch model, the chance probability is assumed to be zero and the slope parameter assumed to be equal for all items. The file labeled “itemdat.LAZ” is used for our example.

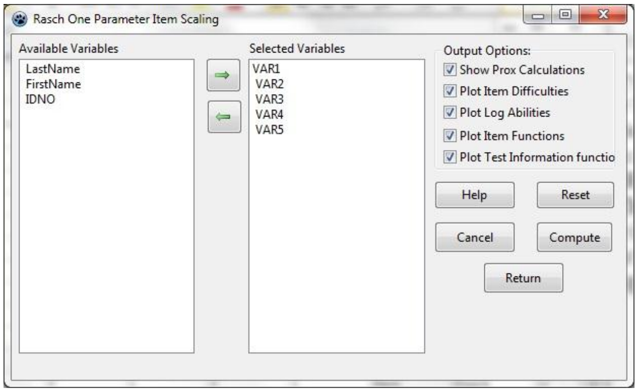


Fig. 11.8 Rasch Item Analysis Dialog

Shown below is a sample of output from a test analyzed by the Rasch model. The model cannot make ability estimates for subjects that miss all items or get all items correct so they are screened out. Parameters estimated are given in log units. Also shown is one of the item information function curve plots. Each item provides the maximum discrimination (information) at that point where the log ability of the subject is approximately the same as the log difficulty of the item. In examining the output you will note that item 1 does not appear to fit the assumptions of the Rasch model as measured by the chi-square statistic.

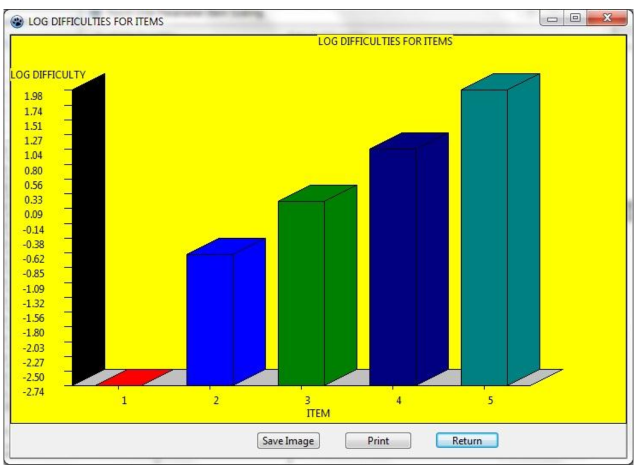


Fig. 11.9 Rasch Item Log Difficulty Estimate Plot

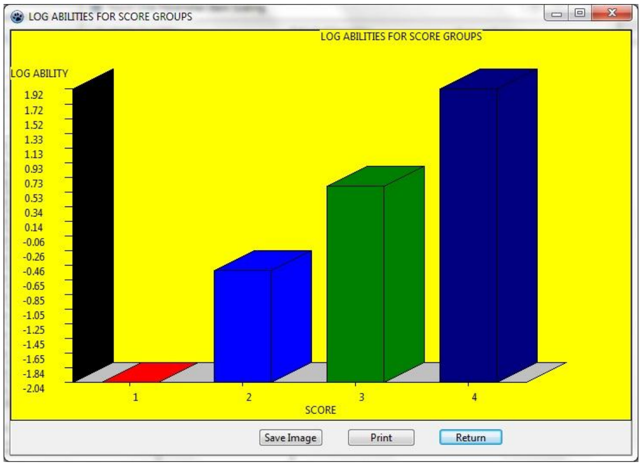


Fig. 11.10 Rasch Log Score Estimates

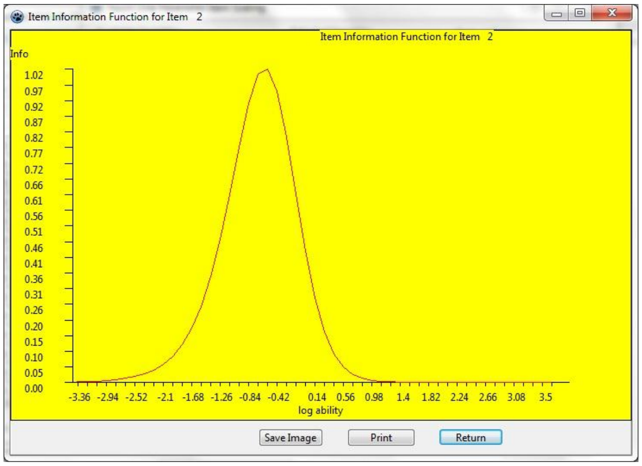


Fig. 11.11 A Rasch Item Characteristic Curve

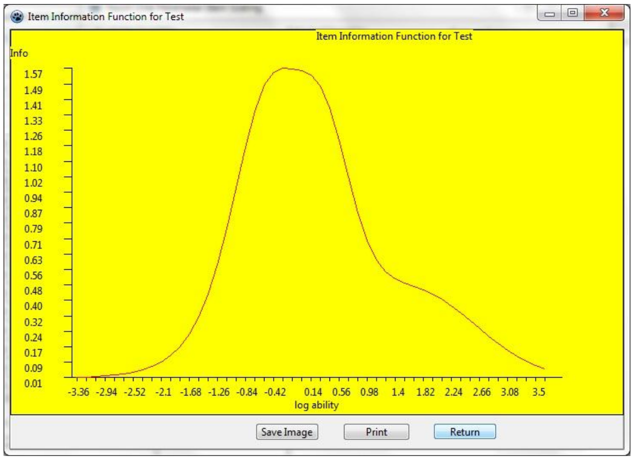


Fig. 11.12 A Rasch Test Information Curve

```
Rasch One-Parameter Logistic Test Scaling (Item Response Theory)
Written by William G. Miller

case    1 eliminated.  Total score was    5
```

```
Case   2 Total Score :=   4 Item scores 1 1 1 1 0
Case   3 Total Score :=   3 Item scores 1 1 1 0 0
Case   4 Total Score :=   2 Item scores 1 1 0 0 0
Case   5 Total Score :=   1 Item scores 1 0 0 0 0
case   6 eliminated.  Total score was   0
Case   7 Total Score :=   1 Item scores 1 0 0 0 0
Case   8 Total Score :=   2 Item scores 1 1 0 0 0
Case   9 Total Score :=   3 Item scores 1 1 1 0 0
Case  10 Total Score :=   4 Item scores 1 1 1 1 0
case  11 eliminated.  Total score was   5
Case  12 Total Score :=   3 Item scores 1 0 1 0 1
Case  13 Total Score :=   3 Item scores 0 1 1 1 0
Case  14 Total Score :=   4 Item scores 1 1 1 0 1
Case  15 Total Score :=   4 Item scores 1 1 0 1 1
Case  16 Total Score :=   2 Item scores 1 0 0 1 0


Total number of score groups :=    4

Matrix of Item Failures in Score Groups
   Score Group   1   2   3   4     Total
ITEM

   1             0   0   1   0      1
   2             2   1   1   0      4
   3             2   3   0   1      6
   4             2   2   3   1      8
   5             2   3   3   2     10
Total            2   3   4   4     13

Item Log Odds Deviation Squared Deviation
  1   -2.48    -2.13       4.54
  2   -0.81    -0.46       0.21
  3   -0.15     0.20       0.04
  4    0.47     0.83       0.68
  5    1.20     1.56       2.43
Score Frequency Log Odds Freq.x Log  Freq.x Log Odds Squared
  1     2       -1.39       -2.77       3.84
  2     3       -0.41       -1.22       0.49
  3     4        0.41        1.62       0.66
  4     4        1.39        5.55       7.69

Prox values and Standard Errors

Item     Scale Value     Standard Error
  1     -2.730           1.334
  2     -0.584           0.770
  3      0.258           0.713
  4      1.058           0.731
  5      1.999           0.844
Y expansion factor :=   1.2821

Score    Scale Value     Standard Error
  1     -1.910          1.540
  2     -0.559          1.258
  3      0.559          1.258
  4      1.910          1.540
X expansion factor =   1.3778
Maximum Likelihood Iteration Number  0
Maximum Likelihood Iteration Number  1
Maximum Likelihood Iteration Number  2
Maximum Likelihood Iteration Number  3

Maximum Likelihood Estimates

Item  Log Difficulty
  1      -2.74
  2      -0.64
  3       0.21
  4       1.04
  5       1.98

Score   Log Ability
  1     -2.04
```

```
  2     -0.54
  3      0.60
  4      1.92

Goodness of Fit Test for Each Item
Item  Chi-Squared  Degrees of  Probability
No.   Value        Freedom     of Larger Value
  1      29.78       9          0.0005
  2       8.06       9          0.5283
  3      10.42       9          0.3177
  4      12.48       9          0.1875
  5       9.00       9          0.4371


Item Data Summary
ITEM  PT.BIS.R.  BIS.R.  SLOPE   PASSED  FAILED  RASCH DIFF
  1   -0.064  -0.117    -0.12   12.00     1      -2.739
  2    0.648   0.850     1.61    9.00     4      -0.644
  3    0.679   0.852     1.63    7.00     6       0.207
  4    0.475   0.605     0.76    5.00     8       1.038
  5    0.469   0.649     0.85    3.00    10       1.981
```

Guttman Scalogram Analysis

Guttman scales are those measurement instruments composed of items which, ideally, form a hierarchy in which the total score of a subject can indicate the actual response (correct or incorrect) of each item. Items are arranged in order of the proportion of subjects passing the item and subjects are grouped and sequenced by their total scores. If the items measure consistently, a triangular pattern should emerge. A coefficient of “reproducibility” is obtained which may be interpreted in a manner similar to test reliability.

Dichotomously scored (0 and 1) items representing the responses of subjects in your data grid rows are the variables (grid columns) analyzed. Select the items to analyze in the same manner as you would for the Classical Item Analysis or the Rasch analysis. When you click the OK button, you will immediately be presented with the results on the output form. An example is shown below.

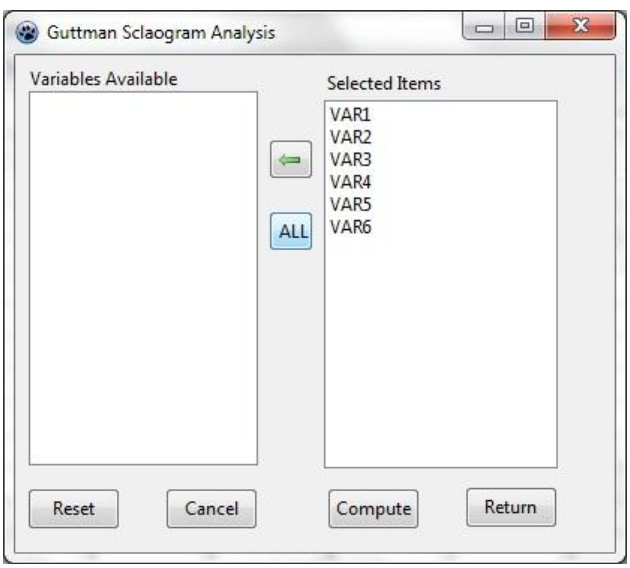


Fig. 11.13 Guttman Scalogram Analysis Dialog

```
GUTTMAN SCALOGRAM ANALYSIS
               Cornell Method
No. of Cases :=  12.  No. of items :=   6
RESPONSE MATRIX
Subject Row           Item Number
Label Sum    VAR1      VAR2      VAR3      VAR4      VAR5      VAR6
          0   1   0   1   0   1   0   1   0   1   0   1
  1   6   0   1   0   1   0   1   0   1   0   1   0   1
  2   6   0   1   0   1   0   1   0   1   0   1   0   1
  3   6   0   1   0   1   0   1   0   1   0   1   0   1
  4   6   0   1   0   1   0   1   0   1   0   1   0   1
  5   6   0   1   0   1   0   1   0   1   0   1   0   1
  6   6   0   1   0   1   0   1   0   1   0   1   0   1
  7   6   0   1   0   1   0   1   0   1   0   1   0   1
  8   6   0   1   0   1   0   1   0   1   0   1   0   1
  9   6   0   1   0   1   0   1   0   1   0   1   0   1
 10   6   0   1   0   1   0   1   0   1   0   1   0   1
 11   6   0   1   0   1   0   1   0   1   0   1   0   1
 12   6   0   1   0   1   0   1   0   1   0   1   0   1
        -cut-   -cut-   -cut-   -cut-   -cut-   -cut-
TOTALS      0  12   0  12   0  12   0  12   0  12   0  12
ERRORS      0   0   0   0   0   0   0   0   0   0   0   0
Coefficient of Reproducibility :=  1.000

GUTTMAN SCALOGRAM ANALYSIS
Goodenough Modification Using Modal Responses
       MODAL ITEM RESPONSES
TOTAL                ITEMS

VAR1      VAR2      VAR3      VAR4      VAR5       VAR6
   6        1         1         1         1          1         1
   5        1         1         1         1          1         1
   4        1         1         1         1          1         1
   3        1         1         1         1          1         1
   2        1         1         1         1          1         1
   1        1         1         1         1          1         1
   0        1         1         1         1          1         1

No. of Cases :=  12.  No. of items :=   6
RESPONSE MATRIX
Subject Row Error                   Item Number
Label  Sum Count     VAR1      VAR2      VAR3      VAR4      VAR5       VAR6
                    0    1    0    1    0    1    0    1    0    1    0    1
   1     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   2     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   3     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   4     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   5     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   6     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   7     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   8     6    0     0    1    0    1    0    1    0    1    0    1    0    1
   9     6    0     0    1    0    1    0    1    0    1    0    1    0    1
  10     6    0     0    1    0    1    0    1    0    1    0    1    0    1
  11     6    0     0    1    0    1    0    1    0    1    0    1    0    1
  12     6    0     0    1    0    1    0    1    0    1    0    1    0    1
TOTALS              0   12    0   12    0   12    0   12    0   12    0   12
PROPORTIONS      0.00 1.00 0.00 1.00 0.00 1.00 0.00 1.00 0.00 1.00 0.00 1.00
Coefficient of Reproducibility :=  1.000
```

Minimal Marginal Reproducibility := 1.000

Successive Interval Scaling

Successive Interval Scaling was developed as an approximation of Thurstone's Paired Comparisons method for estimation of scale values and dispersion of scale values for items designed to measure attitudes. Typically, five to nine categories are used by judges to indicate the degree to which an item expresses an attitude (if a subject agrees with the item) between very negative to very positive. Once scale values are estimated, the items responded to by subjects are scored by obtaining the median scale value of those items to which the subject agrees.

To obtain Successive interval scale values, select that option under the Measurement group in the Analyses menu on the main form. The specifications form below will appear. Select those items (variables) you wish to scale. The data analyzed consists of rows representing judges and columns representing the scale value chosen for an item by a judge. The file labeled "sucsintv.LAZ" is used as an example file.

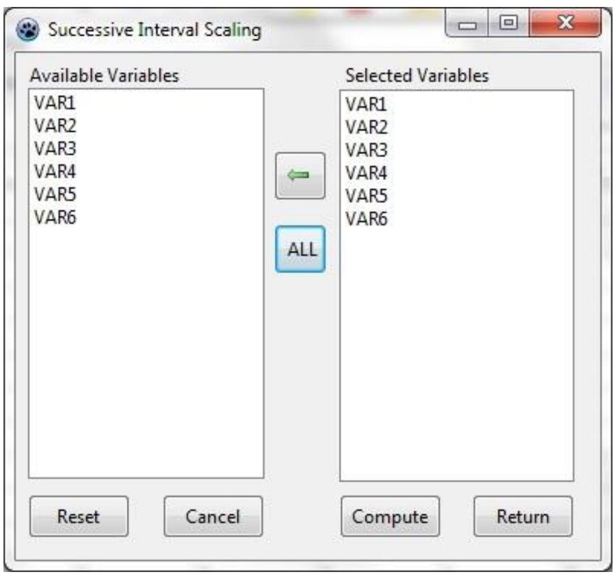


Fig. 11.14 Successive Scaling Dialog

When you click the OK button on the box above, the results will appear on the printout form. An example of results are presented below.

```
          SUCCESSIVE INTERVAL SCALING RESULTS

              0- 1   1- 2   2- 3   3- 4   4- 5   5- 6   6- 7
     VAR1
Frequency        0      0      0      0      4      4      4
Proportion   0.000  0.000  0.000  0.000  0.333  0.333  0.333
Cum. Prop.   0.000  0.000  0.000  0.000  0.333  0.667  1.000
Normal z      -      -      -      -    -0.431  0.431   -
     VAR2
Frequency        0      0      1      3      4      4      0
Proportion   0.000  0.000  0.083  0.250  0.333  0.333  0.000

Cum. Prop.   0.000  0.000  0.083  0.333  0.667  1.000  1.000
Normal z      -      -    -1.383 -0.431  0.431   -      -
     VAR3
Frequency        0      0      4      3      4      1      0
Proportion   0.000  0.000  0.333  0.250  0.333  0.083  0.000
Cum. Prop.   0.000  0.000  0.333  0.583  0.917  1.000  1.000
Normal z      -      -    -0.431  0.210  1.383   -      -
     VAR4
Frequency        0      3      4      5      0      0      0
```

```
Proportion   0.000  0.250  0.333  0.417  0.000  0.000  0.000

Cum. Prop.   0.000  0.250  0.583  1.000  1.000  1.000  1.000
Normal z      -    -0.674  0.210   -      -      -      -
     VAR5
Frequency        5      4      3      0      0      0      0
Proportion   0.417  0.333  0.250  0.000  0.000  0.000  0.000
Cum. Prop.   0.417  0.750  1.000  1.000  1.000  1.000  1.000
Normal z    -0.210  0.674   -      -      -      -      -
      VAR6
Frequency        1      2      2      2      2      2      1
Proportion   0.083  0.167  0.167  0.167  0.167  0.167  0.083

Cum. Prop.   0.083  0.250  0.417  0.583  0.750  0.917  1.000
Normal z    -1.383 -0.674 -0.210  0.210  0.674  1.383   -

                  INTERVAL WIDTHS
            2- 1   3- 2   4- 3   5- 4   6- 5
     VAR1    -      -      -      -     0.861
     VAR2    -      -     0.952  0.861   -
     VAR3    -      -     0.641  1.173   -
     VAR4    -     0.885   -      -      -
     VAR5   0.885   -      -      -      -
      VAR6  0.709  0.464  0.421  0.464  0.709

Mean Width   0.80   0.67   0.67   0.83   0.78
No. Items       2      2      3      3      2
Std. Dev.s   0.02   0.09   0.07   0.13   0.01
Cum. Means   0.80   1.47   2.14   2.98   3.76

ESTIMATES OF SCALE VALUES AND THEIR DISPERSIONS
Item       No. Ratings Scale Value  Discriminal Dispersion
     VAR1       12         3.368       1.224
     VAR2       12         2.559       0.822
     VAR3       12         1.919       0.811
     VAR4       12         1.303       1.192

     VAR5       12         0.199       1.192
      VAR6      12         1.807       0.759

Z scores Estimated from Scale values
             0- 1   1- 2   2- 3   3- 4   4- 5   5- 6   6- 7
     VAR1  -3.368 -2.571 -1.897 -1.225 -0.392  0.392
     VAR2  -2.559 -1.762 -1.088 -0.416  0.416  1.201
     VAR3  -1.919 -1.122 -0.448  0.224  1.057  1.841
     VAR4  -1.303 -0.506  0.169  0.840  1.673  2.458
     VAR5  -0.199  0.598  1.272  1.943  2.776  3.000
      VAR6 -1.807 -1.010 -0.336  0.336  1.168  1.953

Cumulative Theoretical Proportions
             0- 1   1- 2   2- 3   3- 4   4- 5   5- 6   6- 7
     VAR1   0.000  0.005  0.029  0.110  0.347  0.653  1.000
     VAR2   0.005  0.039  0.138  0.339  0.661  0.885  1.000
     VAR3   0.028  0.131  0.327  0.589  0.855  0.967  1.000
     VAR4   0.096  0.306  0.567  0.800  0.953  0.993  1.000
     VAR5   0.421  0.725  0.898  0.974  0.997  0.999  1.000
      VAR6  0.035  0.156  0.369  0.631  0.879  0.975  1.000

Average Discrepency Between Theoretical and Observed Cumulative Proportions =  0.050

Maximum discrepency =  0.200 found in item VAR4
```

Differential Item Functioning

Anyone developing tests today should be sensitive to the fact that some test items may present a bias for one or more subgroups in the population to which the test is administered. For example, because of societal value systems, boys and girls may be exposed to quite different learning experiences during their youth. A word test in mathematics may unintentionally give an advantage to one gender group over another simply by the examples used in the item. To identify possible bias in an item, one can examine the differential item functioning of each item for the sub-groups to which the test is administered. The Mantel-Haenszel test statistic may be applied to test the difference on the item characteristic curve for the difference between a "focus" group and a "reference" group. We will demonstrate using a data set in which 40 items have been administered to 1000 subjects in one group and 1000 subjects in another group. The groups are simply coded 1 and 2 for the reference and focus groups. Since there may be very few (or no) subjects that get a specific total score, we will group the total scores obtained by subjects into groups of 4 so that we are comparing subjects in the groups that have obtained total item scores of 0 to 3, 4 to 7, …, 40 to 43. As you will see, even this grouping is too small for several score groups and we should probably change the score range for the lowest and highest scores to a larger range of scores in another run.

When you elect to do this analysis, the specification form below appears:

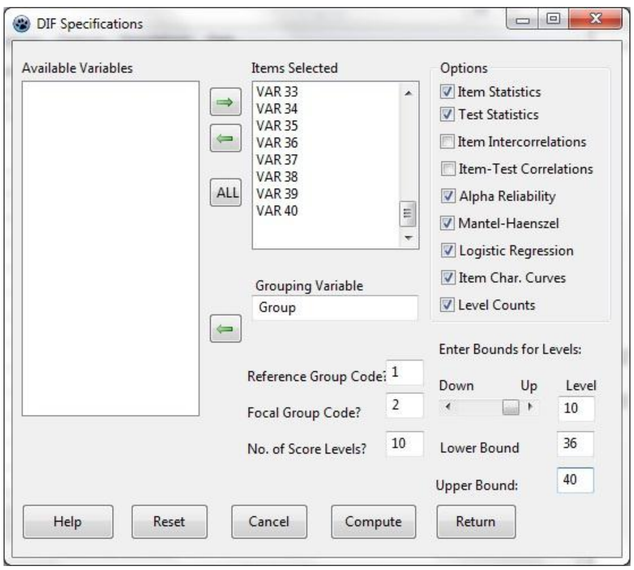


Fig. 11.15 Differential Item Functioning Dialog

On the above form you specify the items to be analyzed and also the variable defining the reference and focus group codes. You may then specify the options desired by clicking the corresponding buttons for the desired options. You also enter the number of score groups to be used in grouping the subject's total scores. When this is specified, you then enter the lowest and highest score for each of those score groups. When you have specified the low and hi score for the first group, click the right arrow on the "slider" bar to move to the next group. You will see that the lowest score has automatically been set to one higher than the previous group's highest score to save you time in entering data. You do not, of course, have to use the same size for the range of each score group. Using too large a range of scores may cut down the sensitivity of the test to differences between the groups. Fairly large samples of subjects is necessary for a reasonable analysis. Once you have completed the specifications, click the Compute button and you will see the following results are obtained (we elected to print the descriptive statistics, correlations and item plots):

```
Mantel-Haenszel DIF Analysis adapted by Bill Miller from
EZDIF written by Niels G. Waller


Total Means with 2000 valid cases.

Variables        VAR 1        VAR 2        VAR 3        VAR 4        VAR 5
```

```
                  0.688        0.064        0.585        0.297        0.451

Variables         VAR 6        VAR 7        VAR 8        VAR 9       VAR 10
                  0.806        0.217        0.827        0.960        0.568

Variables        VAR 11       VAR 12       VAR 13       VAR 14       VAR 15
                  0.350        0.291        0.725        0.069        0.524

Variables        VAR 16       VAR 17       VAR 18       VAR 19       VAR 20
                  0.350        0.943        0.545        0.017        0.985

Variables        VAR 21       VAR 22       VAR 23       VAR 24       VAR 25
                  0.778        0.820        0.315        0.203        0.982

Variables        VAR 26       VAR 27       VAR 28       VAR 29       VAR 30
                  0.834        0.700        0.397        0.305        0.223

Variables        VAR 31       VAR 32       VAR 33       VAR 34       VAR 35
                  0.526        0.585        0.431        0.846        0.115

Variables        VAR 36       VAR 37       VAR 38       VAR 39       VAR 40
                  0.150        0.817        0.909        0.793        0.329


Total Variances with 2000 valid cases.

Variables         VAR 1        VAR 2        VAR 3        VAR 4        VAR 5
                  0.215        0.059        0.243        0.209        0.248

Variables         VAR 6        VAR 7        VAR 8        VAR 9       VAR 10
                  0.156        0.170        0.143        0.038        0.245

Variables        VAR 11       VAR 12       VAR 13       VAR 14       VAR 15
                  0.228        0.206        0.199        0.064        0.250

Variables        VAR 16       VAR 17       VAR 18       VAR 19       VAR 20
                  0.228        0.054        0.248        0.017        0.015

Variables        VAR 21       VAR 22       VAR 23       VAR 24       VAR 25
                  0.173        0.148        0.216        0.162        0.018

Variables        VAR 26       VAR 27       VAR 28       VAR 29       VAR 30
                  0.139        0.210        0.239        0.212        0.173

Variables        VAR 31       VAR 32       VAR 33       VAR 34       VAR 35
                  0.249        0.243        0.245        0.130        0.102

Variables        VAR 36       VAR 37       VAR 38       VAR 39       VAR 40
                  0.128        0.150        0.083        0.164        0.221


Total Standard Deviations with 2000 valid cases.

Variables         VAR 1        VAR 2        VAR 3        VAR 4        VAR 5
                  0.463        0.244        0.493        0.457        0.498

Variables         VAR 6        VAR 7        VAR 8        VAR 9       VAR 10
                  0.395        0.412        0.379        0.196        0.495

Variables        VAR 11       VAR 12       VAR 13       VAR 14       VAR 15
                  0.477        0.454        0.447        0.253        0.500

Variables        VAR 16       VAR 17       VAR 18       VAR 19       VAR 20
                  0.477        0.233        0.498        0.129        0.124

Variables        VAR 21       VAR 22       VAR 23       VAR 24       VAR 25
                  0.416        0.384        0.465        0.403        0.135

Variables        VAR 26       VAR 27       VAR 28       VAR 29       VAR 30
                  0.372        0.459        0.489        0.461        0.416

Variables        VAR 31       VAR 32       VAR 33       VAR 34       VAR 35
                  0.499        0.493        0.495        0.361        0.319
```

```
Variables         VAR 36       VAR 37       VAR 38       VAR 39       VAR 40
                   0.357        0.387        0.288        0.405        0.470

Total Score: Mean =     21.318, Variance =     66.227, Std.Dev. =      8.138

Reference group size = 1000, Focus group size = 1000
Alpha Reliability Estimate for Test = 0.9228  S.E. of Measurement =    2.261
Conditioning Levels
Lower        Upper
    0            3
    4            7
    8           11
   12           15
   16           19
   20           23
   24           27
   28           31
   32           35
   36           40
```

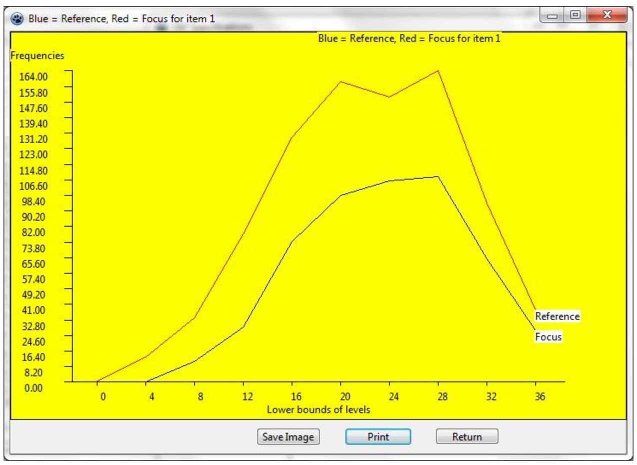


Fig. 11.16 Differential Item Functioning Curve

Etc.

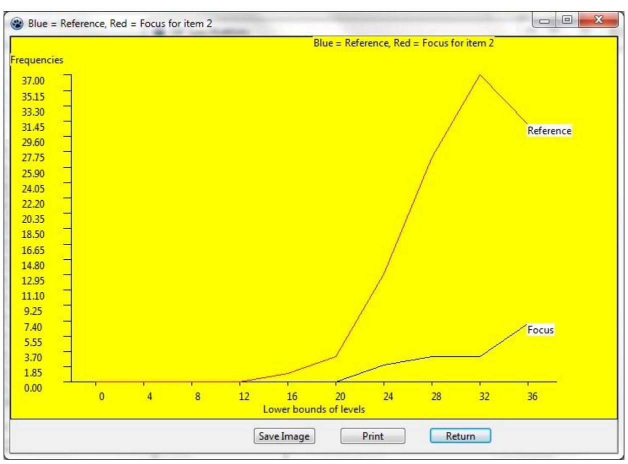


Fig. 11.17 Another Item Differential Functioning Curve

etc. for all items. Note the difference for the two item plots shown above! Next, the output reflects multiple passes to "fit" the data for the M-H test:

COMPUTING M-H CHI-SQUARE, PASS # 1

Cases in Reference Group

Score Level Counts by Item

Variables	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 6	VAR 7	VAR 8	VAR 9	VAR 10
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 11	VAR 12	VAR 13	VAR 14	VAR 15
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 16	VAR 17	VAR 18	VAR 19	VAR 20
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 21	VAR 22	VAR 23	VAR 24	VAR 25
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175

24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 26	VAR 27	VAR 28	VAR 29	VAR 30
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 31	VAR 32	VAR 33	VAR 34	VAR 35
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Score Level Counts by Item

Variables	VAR 36	VAR 37	VAR 38	VAR 39	VAR 40
0- 3	8	8	8	8	8
4- 7	38	38	38	38	38
8- 11	65	65	65	65	65
12- 15	108	108	108	108	108
16- 19	153	153	153	153	153
20- 23	175	175	175	175	175
24- 27	154	154	154	154	154
28- 31	167	167	167	167	167
32- 35	94	94	94	94	94
36- 40	38	38	38	38	38

Cases in Focus Group

Score Level Counts by Item

Variables	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables	VAR 6	VAR 7	VAR 8	VAR 9	VAR 10

0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables

	VAR 11	VAR 12	VAR 13	VAR 14	VAR 15
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables

	VAR 16	VAR 17	VAR 18	VAR 19	VAR 20
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables

	VAR 21	VAR 22	VAR 23	VAR 24	VAR 25
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables

	VAR 26	VAR 27	VAR 28	VAR 29	VAR 30
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables

	VAR 31	VAR 32	VAR 33	VAR 34	VAR 35
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Score Level Counts by Item

Variables

	VAR 36	VAR 37	VAR 38	VAR 39	VAR 40
0- 3	7	7	7	7	7
4- 7	47	47	47	47	47
8- 11	94	94	94	94	94
12- 15	139	139	139	139	139
16- 19	177	177	177	177	177
20- 23	174	174	174	174	174
24- 27	141	141	141	141	141
28- 31	126	126	126	126	126
32- 35	68	68	68	68	68
36- 40	27	27	27	27	27

Insufficient data found in level: 0 - 3

CODES	ITEM	SIG.	ALPHA	CHI2	P-VALUE	MH D-DIF	S.E. MH D-DIF
C R	1	***	8.927	276.392	0.000	-5.144	0.338
C R	2	***	10.450	68.346	0.000	-5.514	0.775
C R	3	***	7.547	280.027	0.000	-4.750	0.305
C R	4	***	10.227	298.341	0.000	-5.464	0.350
C R	5	***	12.765	393.257	0.000	-5.985	0.339
B	6	***	0.571	15.476	0.000	1.316	0.331
A	7	*	0.714	6.216	0.013	0.791	0.310
A	8	*	0.705	5.694	0.017	0.822	0.335
B	9	**	0.493	6.712	0.010	1.664	0.621
B	10	***	0.621	17.349	0.000	1.121	0.267
A	11	*	0.775	4.511	0.034	0.599	0.275
A	12	***	0.687	9.833	0.002	0.883	0.277
B	13	***	0.647	11.904	0.001	1.024	0.294
B	14	**	0.568	7.160	0.007	1.331	0.482
B	15	***	0.600	19.747	0.000	1.199	0.267
B	16	***	0.601	18.326	0.000	1.198	0.278
A	17		0.830	0.486	0.486	0.438	0.538
A	18	***	0.709	8.989	0.003	0.807	0.264
A	19		0.582	1.856	0.173	1.270	0.834
A	20		1.991	1.769	0.183	-1.618	1.072
A	21	*	0.725	5.783	0.016	0.754	0.308
A	22	*	0.743	4.023	0.045	0.697	0.337
B	23	***	0.572	20.804	0.000	1.315	0.286
A	24	*	0.723	5.362	0.021	0.764	0.321
A	25		0.555	1.782	0.182	1.385	0.915
B	26	***	0.540	16.456	0.000	1.447	0.353
A	27	***	0.687	9.240	0.002	0.884	0.287
A	28	**	0.735	6.822	0.009	0.723	0.271
A	29	***	0.681	9.458	0.002	0.904	0.289
A	30	*	0.756	4.342	0.037	0.658	0.306
A	31	***	0.724	8.016	0.005	0.758	0.263
A	32	*	0.745	6.513	0.011	0.693	0.266
A	33	**	0.738	6.907	0.009	0.715	0.267
A	34		0.944	0.089	0.766	0.135	0.360
A	35		0.769	2.381	0.123	0.618	0.383
A	36		0.819	1.530	0.216	0.469	0.357
A	37	*	0.709	5.817	0.016	0.809	0.326
A	38	*	0.665	4.552	0.033	0.960	0.431
A	39		0.779	3.305	0.069	0.588	0.312
B	40	***	0.644	13.215	0.000	1.034	0.280

No. of items purged in pass 1 = 5
Item Numbers:

```
1
2
3
4
5
CODES ITEM     SIG.  ALPHA   CHI2    P-VALUE    MH D-DIF   S.E. MH D-DIF
C R    1      ***  9.367   283.535   0.000     -5.257        0.343
C R    2      ***  8.741    65.854   0.000     -5.095        0.704
C R    3      ***  7.923   287.705   0.000     -4.864        0.310
C R    4      *** 10.888   305.319   0.000     -5.611        0.358
C R    5      *** 13.001   399.009   0.000     -6.028        0.340
B      6      ***  0.587    13.927   0.000      1.251        0.331
A      7        *  0.725     5.598   0.018      0.756        0.311
A      8        *  0.724     4.851   0.028      0.760        0.335
B      9        *  0.506     6.230   0.013      1.599        0.620
B     10      ***  0.638    15.345   0.000      1.056        0.267
A     11           0.798     3.516   0.061      0.529        0.274
A     12      ***  0.700     8.907   0.003      0.838        0.276
A     13      ***  0.663    10.414   0.001      0.964        0.294
B     14        *  0.595     6.413   0.011      1.219        0.466
B     15      ***  0.616    17.707   0.000      1.139        0.268
B     16      ***  0.617    16.524   0.000      1.133        0.276
A     17           0.850     0.355   0.551      0.382        0.537
A     18       **  0.729     7.642   0.006      0.742        0.263
A     19           0.595     1.721   0.190      1.222        0.831
A     20           2.004     1.805   0.179     -1.633        1.073
A     21        *  0.746     4.790   0.029      0.688        0.307
A     22           0.773     2.996   0.083      0.606        0.336
B     23      ***  0.573    20.155   0.000      1.307        0.289
A     24        *  0.736     4.796   0.029      0.722        0.320
A     25           0.570     1.595   0.207      1.320        0.914
B     26      ***  0.554    14.953   0.000      1.388        0.354
A     27       **  0.707     7.819   0.005      0.816        0.287
A     28        *  0.750     5.862   0.015      0.675        0.272
A     29      ***  0.704     7.980   0.005      0.825        0.286
A     30        *  0.769     3.845   0.050      0.618        0.305
A     31       **  0.743     6.730   0.009      0.698        0.263
A     32        *  0.762     5.551   0.018      0.640        0.266
A     33        *  0.749     6.193   0.013      0.681        0.268
A     34           0.976     0.007   1.000      0.058        0.360
A     35           0.790     1.975   0.160      0.555        0.375
A     36           0.832     1.310   0.252      0.432        0.354
A     37        *  0.721     5.148   0.023      0.770        0.329
A     38        *  0.678     4.062   0.044      0.914        0.433
A     39           0.804     2.490   0.115      0.512        0.312
A     40      ***  0.664    11.542   0.001      0.963        0.279

No. of items purged in pass 1 = 5
Item Numbers:
1
2
3
4
5
```

One should probably combine the first two score groups (0-3 and 4-7) into one group and the last three groups into one group so that sufficient sample size is available for the comparisons of the two groups. This would, of course, reduce the number of groups from 11 in our original specifications to 8 score groups. The chi-square statistic identifies items you will want to give specific attention. Examine the data plots for those items. Differences found may suggest bias in those items. Only examination of the actual content can help in this decision. Even though two groups may differ in their item response patterns does not provide sufficient grounds to establish bias - perhaps it simply identifies a true difference in educational achievement due to other factors.

Adjustment of Reliability For Variance Change

Researchers will sometimes use a test that has been standardized on a large, heterogenous population of subjects. Such tests typically report rather high internal-consistency reliability estimates (e.g. Cronbach's estimate.)

But what is the reliability if one administers the test to a much more homogeneous population? For example, assume a high school counselor administers a "College Aptitude Test" that reports a reliability of 0.95 with a standard deviation of 15 (variance of 225) and a mean of 20.0 for the national norm. What reliability would the counselor expect to obtain for her sample of students that obtain a mean of 22.8 and a standard deviation of 10.2 (variance of 104.04)? This procedure will help provide the estimate. Shown below is the specification form and our sample values entered. When the compute button is clicked, the results shown are obtained.

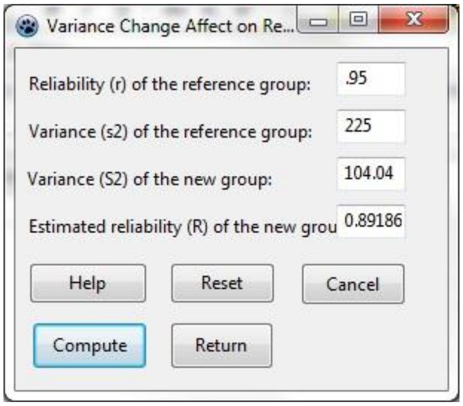


Fig. 11.18 Reliability Adjustment for Variability Dialog

Polytomous DIF Analysis

The purpose of the differential item functioning programs are to identify test or attitude items that "perform" differently for two groups - a target group and a reference group. Two procedures are provided and selected on the basis of whether the items are dichotomous (0 and 1 scoring) or consist of multiple categories (e.g. Likert responses ranging from 1 to 5.) The latter case is where the Polytomous DIF Analysis is selected. When you initiate this procedure you will see the dialogue box shown below:

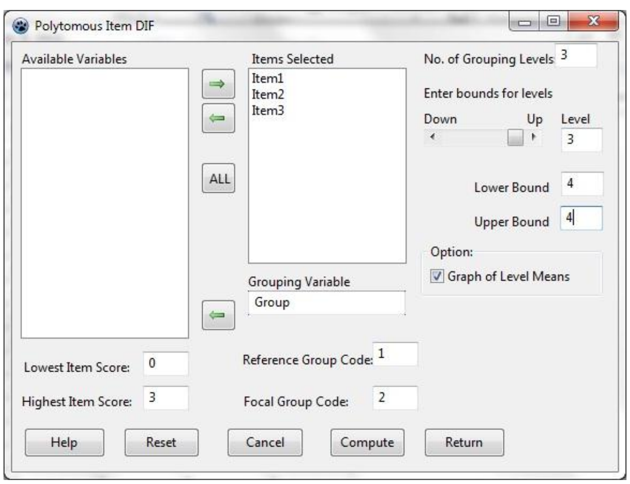


Fig. 11.19 Polytomous Item Differential Item Functioning Dialog

The results from an analysis of three items with five categories that have been collapsed into three category levels is shown below. A sample of 500 subject's attitude scores were observed.

```
Polytomous Item DIF Analysis adapted by Bill Miller from
Procedures for extending item bias detection techniques
by Catherine Welch and H.D. Hoover, 1993
Applied Measurement in Education 6(1), pages 1-19.

Conditioning Levels
Lower        Upper
```

```
    0            1
    2            3
    4            4

Observed Category Frequencies
Item  Group  Level  Category Number
                       0         1         2         3
  1   Ref.    1        0        55        56        40
  1   Focal   1        0        50        47        38
  1   Total   1        0       105       103        78

  1   Ref.    2        0         2         0         0
  1   Focal   2        0         1         0         0
  1   Total   2        0         3         0         0

  1   Ref.    3        0         3         3         0
  1   Focal   3        0         5         0         0
  1   Total   3        0         8         3         0

t-test values for Reference and Focus Means for each level
Mean Reference =      1.901 SD =     14.006 N =   151
Mean Focal     =      1.911 SD =     13.338 N =   135
Level   1 t =    0.006 with deg. freedom =   284
Mean Reference =      2.000 SD =      2.000 N =     2
Mean Focal     =      1.000 SD =      1.000 N =     1
Level   2 t =    0.000 with deg. freedom =     0
Mean Reference =      1.500 SD =      2.510 N =     6
Mean Focal     =      1.000 SD =      2.236 N =     5
Level   3 t =   -0.313 with deg. freedom =     9
Composite z statistic = -0.134. Prob. > |z| =  0.553
Weighted Composite z statistic = -0.030. Prob. > |z| =  0.512
Generalized Mantel-Haenszel =      0.008 with D.F. = 1 and Prob. > Chi-Sqr. =  1.000
```

Fig. 11.20 Level Means for Polytomous Item

For Item 2:

```
Observed Category Frequencies
Item  Group  Level  Category Number
                       0         1         2         3
  2   Ref.    1        0        63        53        48
  2   Focal   1        0        44        46        50
  2   Total   1        0       107        99        98

  2   Ref.    2        0         2         0         0
  2   Focal   2        0         1         0         0
  2   Total   2        0         3         0         0

  2   Ref.    3        0         5         1         0
  2   Focal   3        0         2         3         0
  2   Total   3        0         7         4         0

t-test values for Reference and Focus Means for each level
Mean Reference =      1.909 SD =     14.725 N =   164
Mean Focal     =      2.043 SD =     15.248 N =   140
```

```
Level   1 t =    0.078 with deg. freedom =   302
Mean Reference =       2.000 SD =       2.000 N =     2
Mean Focal     =       1.000 SD =       1.000 N =     1
Level   2 t =    0.000 with deg. freedom =     0
Mean Reference =       1.167 SD =       2.041 N =     6
Mean Focal     =       1.600 SD =       2.608 N =     5
Level   3 t =    0.279 with deg. freedom =     9
Composite z statistic =  0.156. Prob. > |z| =  0.438
Weighted Composite z statistic =  0.735. Prob. > |z| =  0.231
Generalized Mantel-Haenszel =      2.480 with D.F. = 1 and Prob. > Chi-Sqr. =  0.115
```

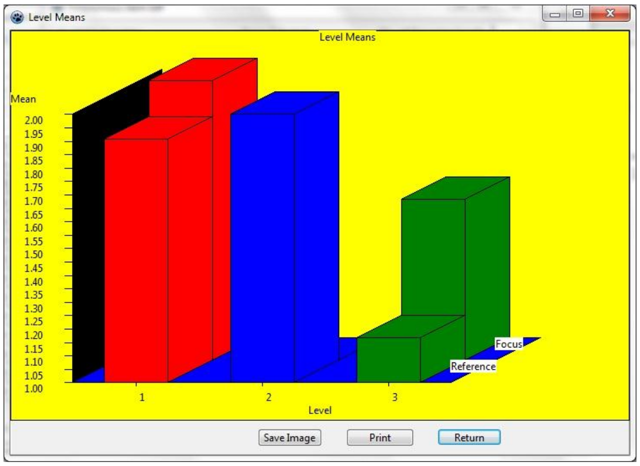


Fig. 11.21 Level Means

```
Observed Category Frequencies
Item   Group   Level   Category Number
                          0          1          2          3
  3    Ref.     1         0         39         46         55
  3    Focal    1         0         46         49         41
  3    Total    1         0         85         95         96

  3    Ref.     2         0          2          0          0
  3    Focal    2         0          1          0          0
  3    Total    2         0          3          0          0

  3    Ref.     3         0          4          2          0
  3    Focal    3         0          3          2          0
  3    Total    3         0          7          4          0
t-test values for Reference and Focus Means for each level
Mean Reference =       2.114 SD =      16.223 N =    140
Mean Focal     =       1.963 SD =      13.964 N =    136
Level   1 t =   -0.082 with deg. freedom =   274
Mean Reference =       2.000 SD =       2.000 N =     2
Mean Focal     =       1.000 SD =       1.000 N =     1
Level   2 t =    0.000 with deg. freedom =     0
Mean Reference =       1.333 SD =       2.066 N =     6
Mean Focal     =       1.400 SD =       1.949 N =     5
Level   3 t =    0.049 with deg. freedom =     9
Composite z statistic = -0.014. Prob. > |z| =  0.506
Weighted Composite z statistic = -0.657. Prob. > |z| =  0.744
Generalized Mantel-Haenszel =      2.279 with D.F. = 1 and Prob. > Chi-Sqr. =
0.131
```

Generate Test Data

To help you become familiar with some of the measurement procedures, you can experiment by creating “artificial” item responses to a test. When you select the option to generate simulated test data, you complete the information in the following specification form. An example is shown. Before you begin, be sure you have closed any open file already in the data grid since the data that is generated will be placed in that grid.

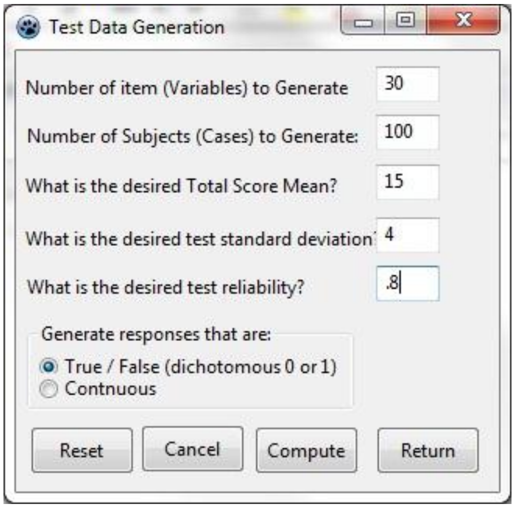
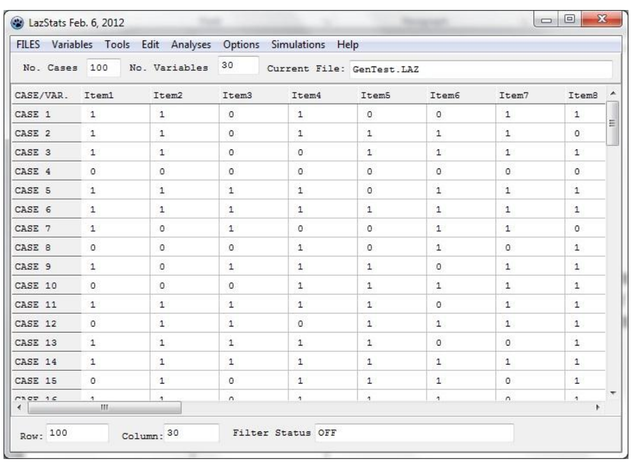


Fig. 11.22 Test Item Generation Dialog

Shown below is a “snap-shot” of the generated test item responses. An additional row has been inserted for the first case which consists of all 1’s. It will serve as the “correct” response for scoring each of the item responses of the subsequent cases. You can save your generated file for future analyses or other work.



CASE/VAR.	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8
CASE 1	1	1	0	1	0	0	1	1
CASE 2	1	1	0	1	1	1	1	0
CASE 3	1	1	0	0	1	1	1	1
CASE 4	0	0	0	0	0	0	0	0
CASE 5	1	1	1	1	0	1	1	1
CASE 6	1	1	1	1	1	1	1	1
CASE 7	1	0	1	0	0	1	1	0
CASE 8	0	0	0	1	0	1	0	1
CASE 9	1	0	1	1	1	0	1	1
CASE 10	0	0	0	1	1	1	1	1
CASE 11	1	1	1	1	1	0	1	1
CASE 12	0	1	1	0	1	1	1	1
CASE 13	1	1	1	1	1	0	0	1
CASE 14	1	1	1	1	1	1	1	1
CASE 15	0	1	0	1	1	1	0	1

Fig. 11.23 Generated Item Data in the Main Grid

Notice that in our example we specified the creation of test data that would have a reliability of 0.8 for 30 items administered to 100 students. If we analyze this data with our Classical Test Analysis procedure, we obtain the following output:

Alpha Reliability Estimate for Test = 0.9100 S.E. of Measurement = 2.365

Clearly, the test generated from our population specifications yielded a somewhat higher reliability than the 0.8 specified for the reliability. Have we learned something about sampling variability? If you request that the total be placed in the data grid when you use analyze the test, you can also use the descriptive statistics procedure to obtain the sample mean, etc. as shown below:

DISTRIBUTION PARAMETER ESTIMATES

TOTAL (N = 100) Sum = 1597.000
Mean = 15.970 Variance = 62.151 Std.Dev. = 7.884
Std.Error of Mean = 0.788
95.00 percent Confidence Interval for the mean = 14.406 to 17.534
Range = 29.000 Minimum = 1.000 Maximum = 30.000
Skewness = -0.235 Std. Error of Skew = 0.241
Kurtosis = -1.008 Std. Error Kurtosis = 0.478

First Quartile = 9.000
Median = 18.000
Third Quartile = 22.000
Interquartile range = 13.000

The frequencies procedure can plot the total score distribution of our sample with the normal curve as a reference to obtain:

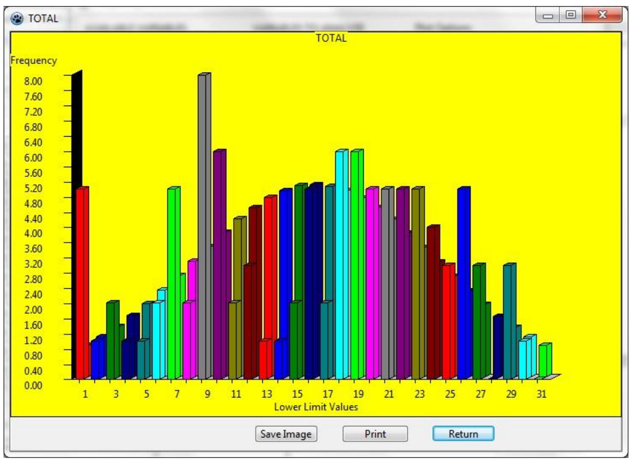


Fig. 11.24 Plot of Generated Test Data

A test of normality of the total scores suggests a possibility that the obtained scores are not normally distributed as shown in the normality test form below:

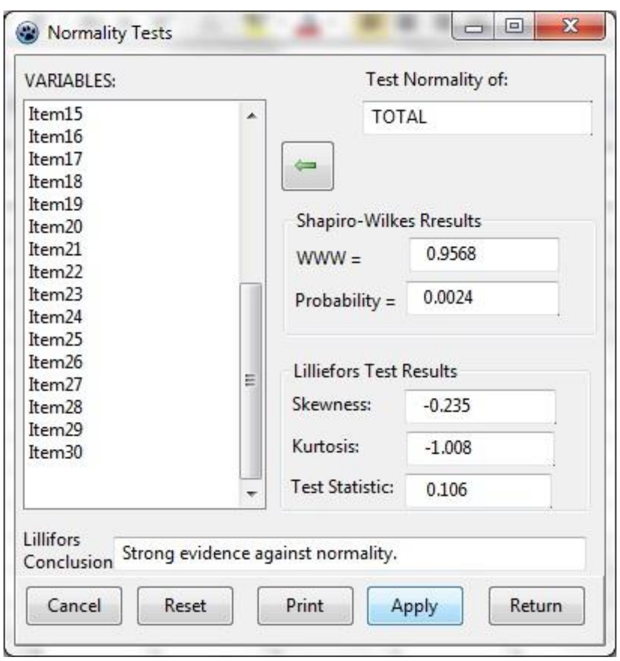


Fig. 11.25 Test of Normality for Generated Data

Spearman-Brown Reliability Prophecy

The Spearman-Brown "Prophecy" formula has been a corner-stone of many instructional text books in measurement theory. Based on "Classical True-Score" theory, it provides an estimate of what the reliability of a test of a different length would be based on the initial test's reliability estimate. It assumes the average difficulty and inter-item covariances of the extended (or trimmed) test are the same as the original test. If these assumptions are valid, it is a reasonable estimator. Shown below is the specification form which appears when you elect this Measurement option from the Analyses menu:

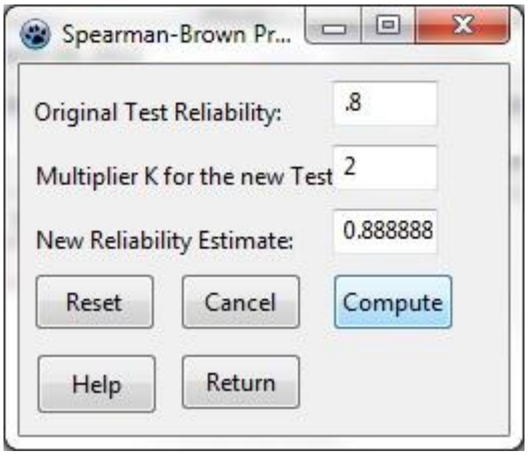


Fig. 11.26 Spearman-Brown Prophecy Dialog

You can see that in an example, that when a test with an initial reliability of 0.8 is doubled (the multiplier k = 2) that the new test is expected to have a reliability of 0.89 approximately. The program may be useful for reducing a test (perhaps by randomly selecting items to delete) that requires too long to administer and has an initially high internal consistency reliability estimate. For example, assume a test of 200 items has a reliability of .95. What is the estimate if the test is reduced by one-half? If the new reliability of 0.9 is satisfactory, considerable time and money may be saved!

Course Grades System

The grade book system is designed for teachers. The teacher can enter information for each student in a course and their raw scores obtained on each test administered during the term. Procedures are built into the grade book system to automatically convert raw test scores into standard (z) scores, ranks and letter grades. When the Grade Book System is first started, you will see the following screen:

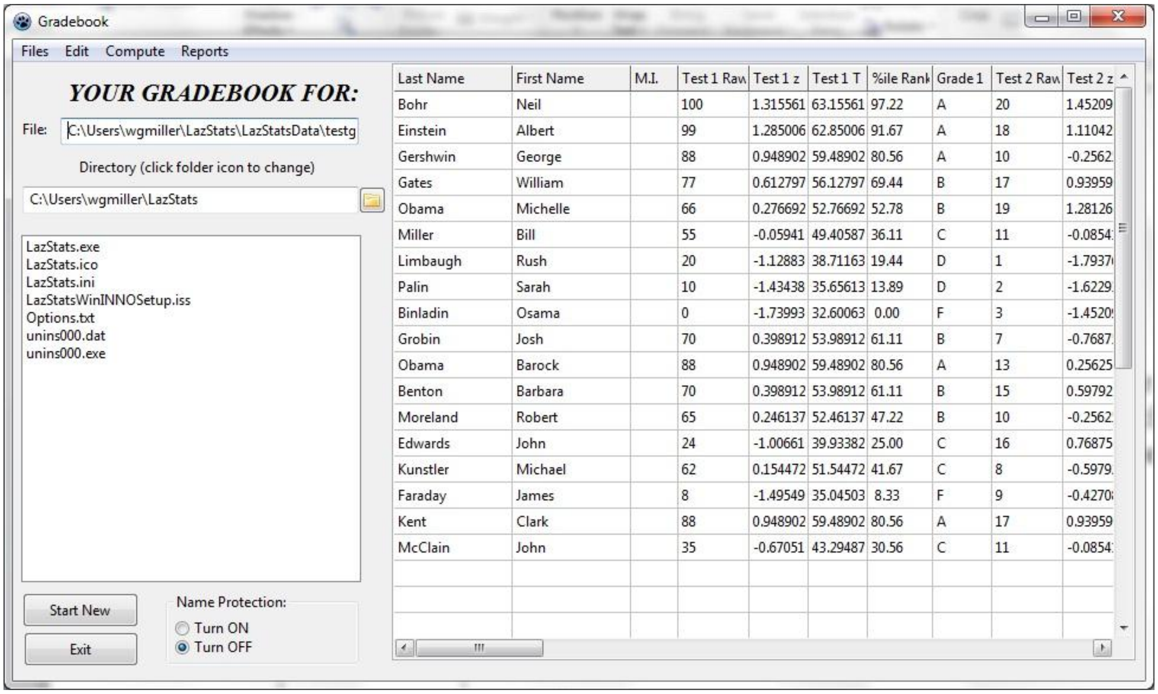


Fig. 11.27 Grading System Dialog with an Opened Grade Book

The teacher can click on the Files menu and create a new grade book after entering name information and the first test scores. When the teacher clicks on the compute button to specify grading a test, the following dialog box appears in which information about the test is recorded:

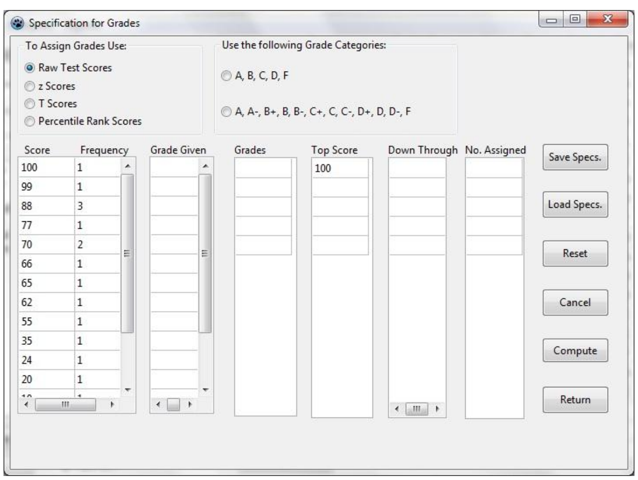


Fig. 11.28 Grading System Test Specification Dialog

To specify the assignment of grades, the top score for each grade interval is entered and the enter key pressed. Clicking the compute button on this form automatically generates z, T and percentile ranks for the students. Upon completion, the following summary is given:

```
Test Analysis Results
Mean =     67.83, Variance =  804.567, Std.Dev. =   28.365

Kuder-Richardson Formula 21 Reliability Estimate = 0.9827

PERCENTILE RANKS
Score Value   Frequency   Cum.Freq. Percentile Rank
___________   __________  __________ ______________
    23.000       1.00       1.00       8.33
```

55.000	1.00	2.00	25.00
56.000	1.00	3.00	41.67
86.000	1.00	4.00	58.33
88.000	1.00	5.00	75.00
99.000	1.00	6.00	91.67

When a test is added, the grid is expanded to include columns for raw scores, z scores, ranks and letter grade. Clicking the Add Student button adds a new row for student information. Students may be added or deleted at any time. Simply place the cursor in the row where a student is to be deleted or added. One can also alphabetize the grid on the student last names. If students are added or deleted, the user can recalculate the z scores, ranks and grades if desired.

The teacher can also request a class report or individual student reports. Shown below is a class report for the sample data above.

```
Class Report
Report for: Bill G Miller

TEST      RAW       Z         T          PERCENTILE GRADE
 NO.      SCORE     SCORE     SCORE      RANK
  1           86    0.640     56.400     58.330     B


Report for: Barb L Benton

TEST      RAW       Z         T          PERCENTILE GRADE
 NO.      SCORE     SCORE     SCORE      RANK
  1           88    0.711     57.100     75.000     B


Report for: Clark A Kent

TEST      RAW       Z         T          PERCENTILE GRADE
 NO.      SCORE     SCORE     SCORE      RANK
  1           56   -0.417     45.800     41.670     F


Report for: Michelle C Obama

TEST      RAW       Z         T          PERCENTILE GRADE
 NO.      SCORE     SCORE     SCORE      RANK
  1           99    1.099     61.000     91.670     A


Report for: George H Bush

TEST      RAW       Z         T          PERCENTILE GRADE
 NO.      SCORE     SCORE     SCORE      RANK
  1           23   -1.581     34.200      8.330     F


Report for: Bill E Clinton

TEST      RAW       Z         T          PERCENTILE GRADE
 NO.      SCORE     SCORE     SCORE      RANK
  1           55   -0.452     45.500     25.000     F
```

Item Banking

The item banking procedure allows the user to explore the concepts of creating a bank of test items. The items stored can then be used to create a printed test to administer to students. A variety of item types may be stored and subsequently selected for a test based on a classification scheme for the content of the items. The type of items that can be included are true-false, multiple choice, matching and essay (or short answer.)

To use the item banking procedure, go to the Measurement option in LazStats and click on the Item Banking option. A sample item bank labeled "testitembank.BNK" has been created and opened with the following dialogue which appears:

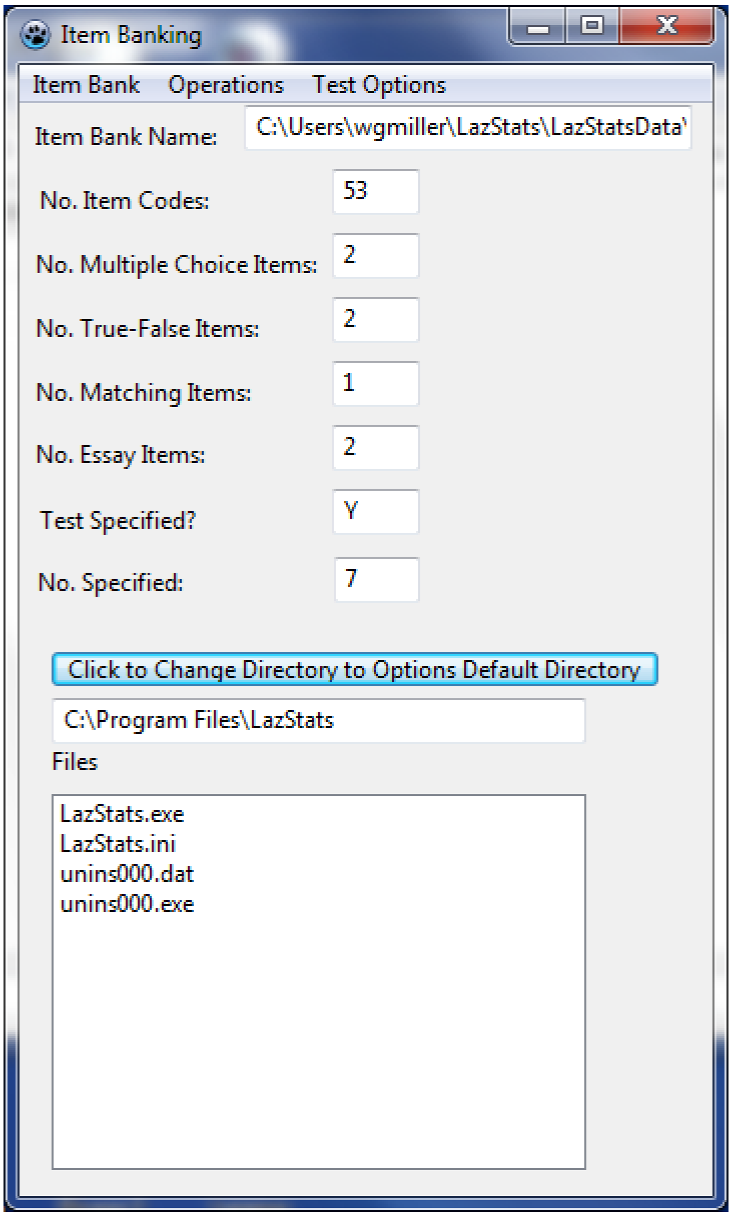


Fig. 11.29 Item Banking Dialogue

Notice that when you open an item bank that you have created, it will be saved with the extension "BNK" and not the usual "TEX" used for a data file. You will have to drop down the type of file from the usual "*.TEX" to all files (*.*) to see the files with the .BNK extension. These files contain a variety of information regarding your item bank including the items themselves, the classification codes for the contents of items, any test specification, etc.. Notice that once you have opened or created a data file, there are three major "drop-down" menus at the top of the item banking dialogue. The "Item Bank" menu includes options to open an item bank, create a new item bank, save an item bank or exit the procedure. The "Operations" menu lets you display all items, create a coding scheme for the items, or create items of a specific type. The "Test Options" menu lets you specify the items for a test, list items for a test or print the test that was specified.

To demonstrate some of these options, the information below was obtained by selecting the Operations menu item to display all item codes:

==

Current Item Coding Scheme

Code number 1
Major Code 1
Minor Code 0
Description Descriptive Statistics

Code number 2
Major Code 1
Minor Code 1
Description Mean

Code number 3
Major Code 1
Minor Code 2
Description Median

Code number 4
Major Code 1
Minor Code 3
Description Mode

Code number 5
Major Code 1
Minor Code 4
Description Variance
. . . (etc.)
Code number 52
Major Code 7
Minor Code 4
Description Attitude Measurement

Code number 53
Major Code 7
Minor Code 5
Description Composite Test Reliability

Notice that each code has a "Major Code" and a "Minor Code" followed by a short description of the corresponding content of the item using that code.

Now we will examine a sample test specification using the "Test Options" item to list all items:

==

Item number: 1
Major code: 7
Minor code: 2
Item type: MC
Item number: 2
Major code: 1
Minor code: 0
Item type: MC
Item number: 1
Major code: 1
Minor code: 0
Item type: TF
Item number: 2
Major code: 1
Minor code: 1
Item type: TF
Item number: 1
Major code: 1
Minor code: 0
Item type: Essay
Item number: 2
Major code: 1
Minor code: 1
Item type: Essay
Item number: 1
Major code: 1
Minor code: 0
Item type: Matching

Notice that the listing only specifies an item number, Major and Minor codes and an item type. This is the list that is used to actually print a test. Using the above specifications for a test, we will now select to print a test and show the first page of the generated test:

==

Directions: This test may contain a variety of different item types.
For each item, circle the correct answer or provide the answer if
required. You may use the back of the test to provide answers to
essay questions - just start with the item number.
Start now!

MULTIPLE CHOICE ITEMS:
Item 1
Which of the following statements is false?
A. Interval measured data can be correctly analyzed with parametric procedures.

B. Ordinal measured data are analyzed using non-parametric procedures.
C. Nominal measured data can be analyzed with parametric procedures.

Item 2
Reference picture = C:\Users\wgmiller\LazStats\HelpFiles\Type1&2Plot.jpg
Type I error is
A. Incorrect sampling.
B. Incorrect research design.
C. Inadequate measurement scale type.
D. Probability of rejecting a true null hypothesis by random sampling variability.
E. Probability of accepting a false null hypothesis by random sampling variability.

TRUE OR FALSE ITEMS:
Item 3
In a highly skewed distribution, the best indicator of central tendency is the mean.
A. TRUE
B. False
Item 4
The sum of squared X values is the same as the square of the sum of X values.
A. TRUE
B. False

ESSAY ITEMS:
Item 5
Expand and simplify the expression sum of (X - Mean)^2
Item 6
Describe kurtosis, mesokurtic, platykurtic

MATCHING ITEMS:
Item 7
A. Normal   1. Monte Carlo Distributions
B. Poisson   2. Theoretical distributions
C. Binomial
D. Beta
E. Gamma

You can alter the printing on the output page before actually printing the test if you desire.

Now we will display the form used to create a new true or false item. We will select the option to "Create or Edit True-False Items" under the "Options" menu:

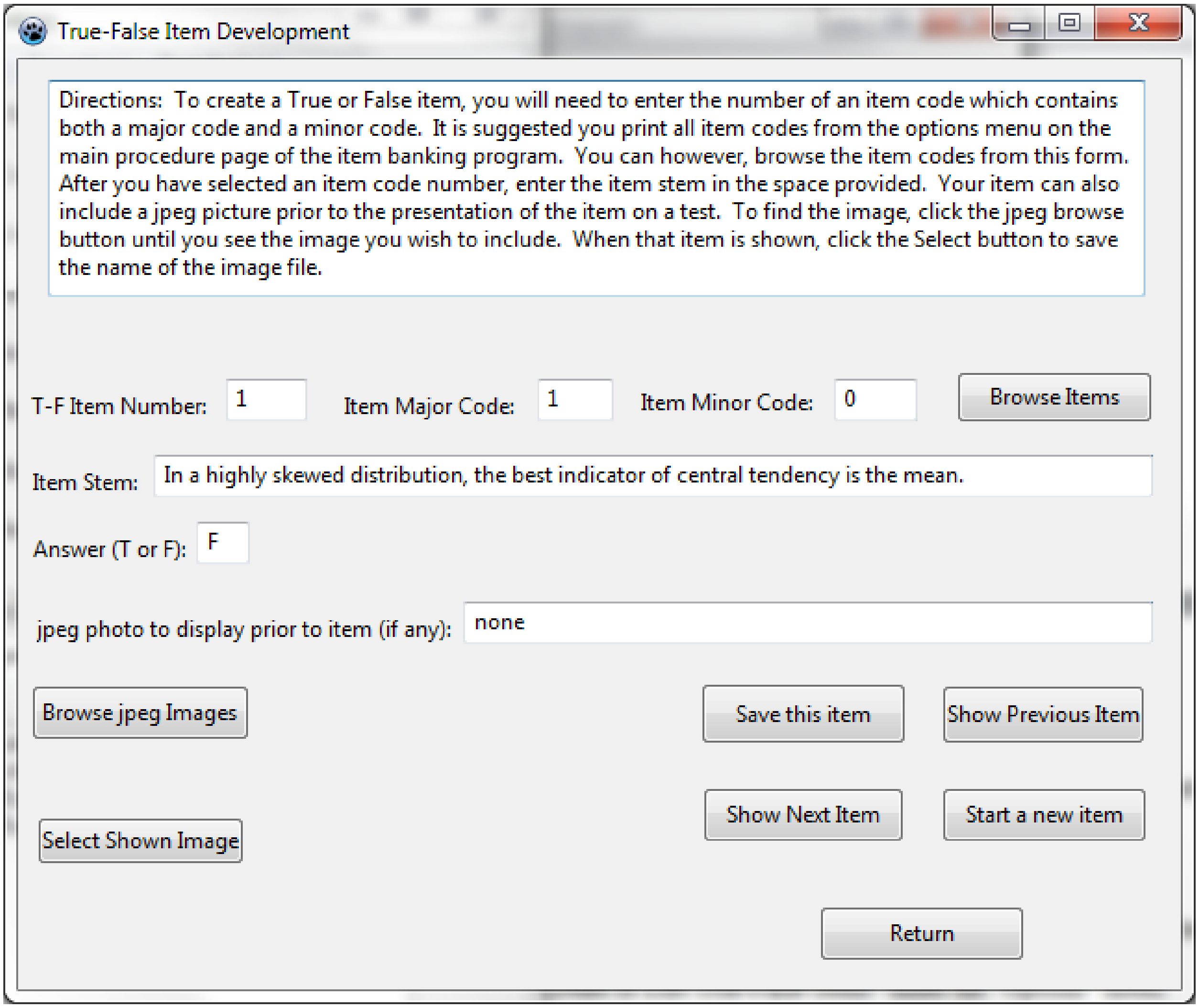


Fig. 11.30 The True or False Item Editing or Creation Dialogue

In the form above you will notice the first previously saved true or false item is displayed. This permits you to edit that item if desired and save it with the "Save this item" button. If you click the "Browse Items" button you will see all items of this type:
Current Items

Item number 1
Major Code 1
Minor Code 0
Item Stem In a highly skewed distribution, the best indicator of central tendency is the mean.
Correct Choice F
Graphic Image none

Item number 2
Major Code 1
Minor Code 1
Item Stem The sum of squared X values is the same as the square of the sum of X values.
Correct Choice F

Graphic Image none

Clicking the "Show Next Item" will take you to the next item previously saved or you can click the "Start a New Item" option to enter a new item and then save it.

Also on each item creation menu there is an option to associate a ".jpg" graphic file with the item. When you print a test, a reference to that item is printed as part of the item (See the first item in the above sample test.) If you save the output of the printed test in a file, you can later open that file in your word processor and insert the graphic image where the reference occurs.

We encourage you to "play" with this procedure to explore the capabilities of an item banking process!

Chapter 12.   Statistical Process Control

XBAR Chart

An Example

We will use the file labeled boltsize.txt to demonstrate the XBAR Chart procedure. Load the file and select the option Statistics / Statistical Process Control / Control Charts / XBAR Chart from the menu. The file contains two variables, lot number and bolt length. These values have been entered in the specification form which is shown below. Notice that the form also provides the option to enter and use a specific "target" value for the process as well as specification levels which may have been provided as guidelines for determining whether or not the process was in control for a given sample.

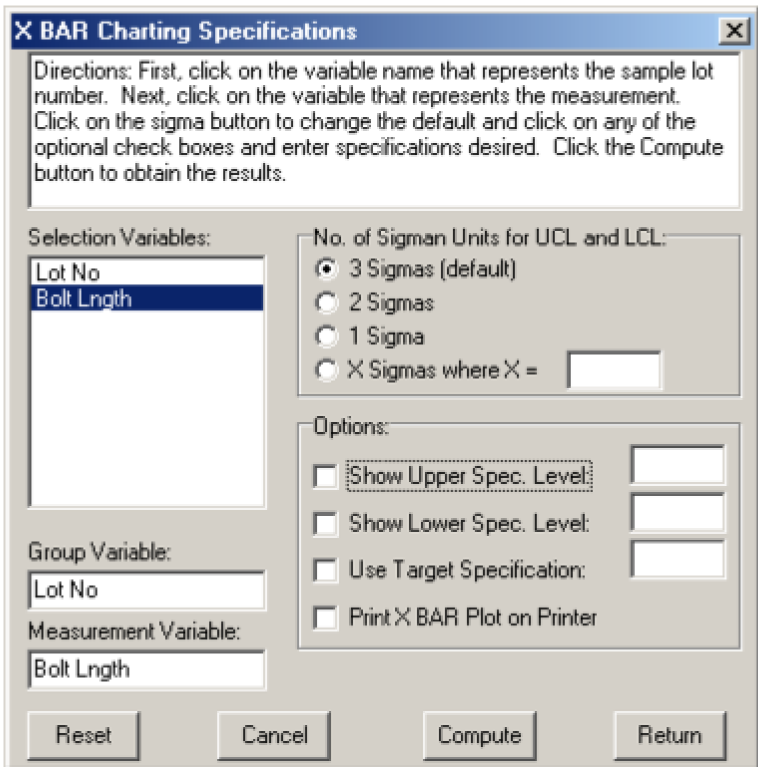


Fig. 12.1 XBAR Chart Dialog

Pressing the Compute button results in the following:

```
X Bar Chart Results

Group Size Mean      Std.Dev.
_____ ____ _________ __________
    1   5    19.88     0.37
    2   5    19.90     0.29
    3   5    20.16     0.27
    4   5    20.08     0.29
    5   5    19.88     0.49
    6   5    19.90     0.39
    7   5    20.02     0.47
    8   5    19.98     0.43
```

```
Grand Mean =    19.97, Std.Dev. =    0.359, Standard Error of Mean =
0.06
Lower Control Limit =   19.805, Upper Control Limit =   20.145
```

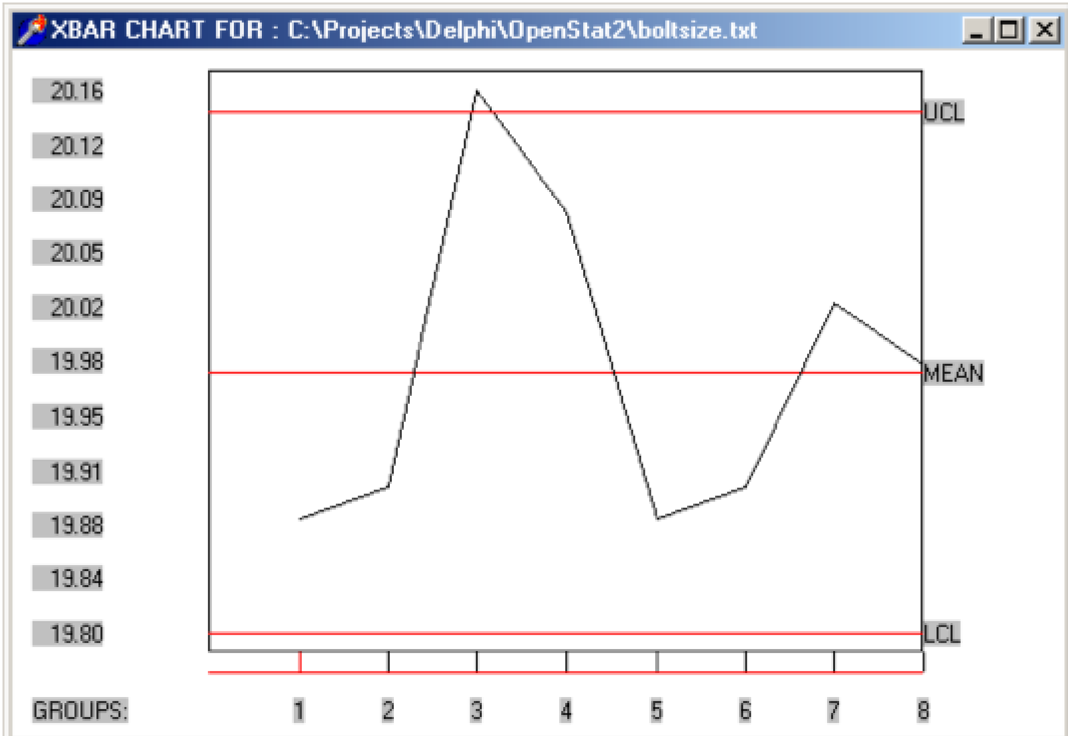


Fig. 12.2 XBAR Chart for Boltsize

If, in addition, we specify a target value of 20 for our bolt and upper and lower specification levels (tolerance) of 20.1 and 19.9, we would obtain the chart shown below:

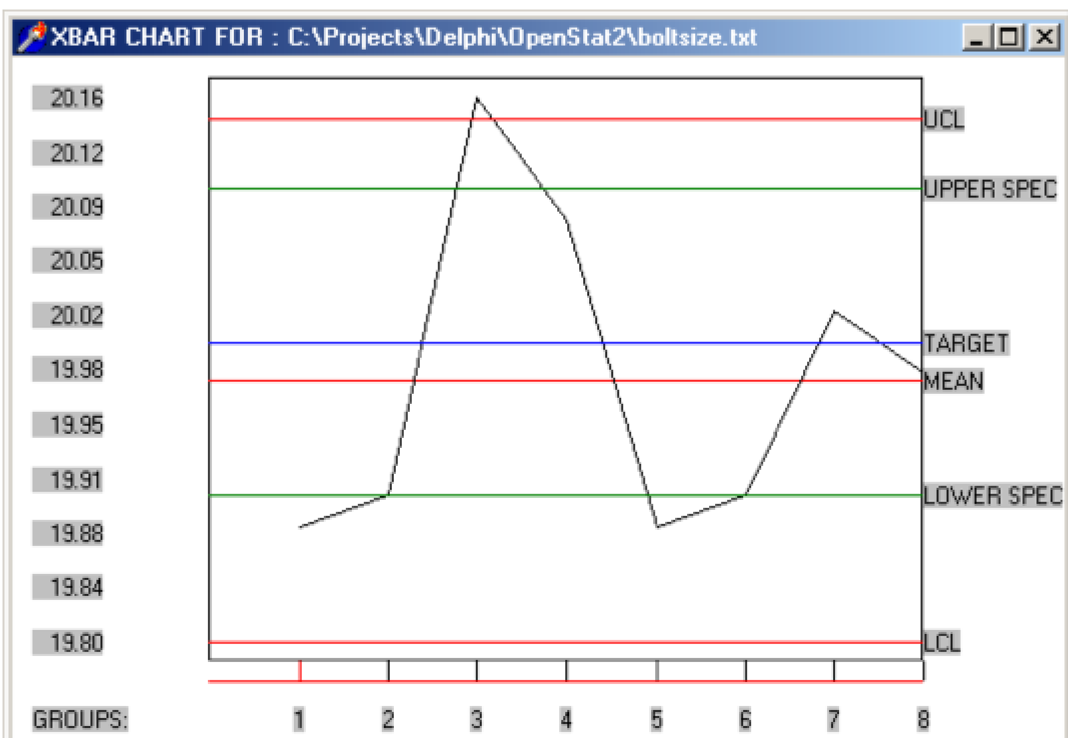


Fig. 12.3 XBAR Chart Plot with Target Specifications

In this chart we can see that the mean of the samples falls slightly below the specified target value and that samples 3 and 5 appear to have bolts outside the tolerance specifications.

Range Chart

As tools wear the products produced may begin to vary more and more widely around the values specified for them. The mean of a sample may still be close to the specified value but the range of values observed may increase. The result is that more and more parts produced may be under or over the specified value. Therefore quality assurance personnel examine not only the mean (XBAR chart) but also the range of values in their sample lots. Again, examine the boltsize.txt file with the option Statistics / Statistical Process Control / Control Charts / Range Chart. Shown below is the specification form and the results:

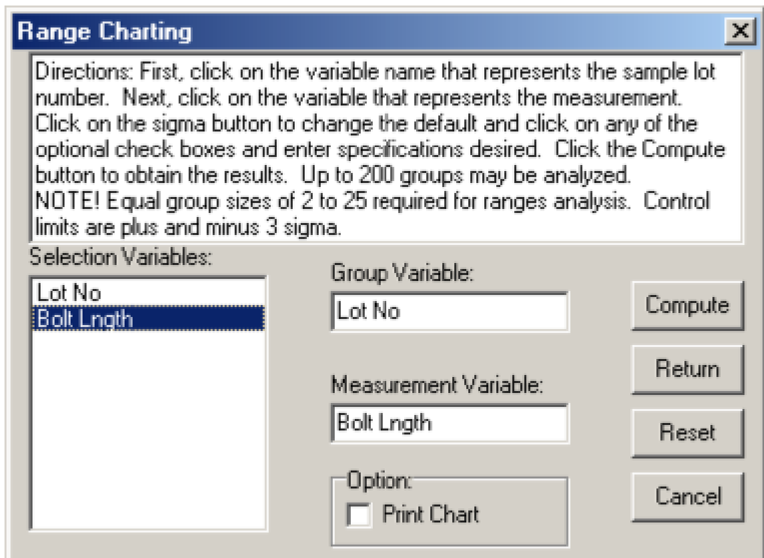


Fig.12.4 Range Chart Dialog

```
X Bar Chart Results

Group Size Mean       Range    Std.Dev.

_____ ____ _________ _______ ________
   1    5    19.88     0.90     0.37
   2    5    19.90     0.70     0.29
   3    5    20.16     0.60     0.27
   4    5    20.08     0.70     0.29
   5    5    19.88     1.20     0.49
   6    5    19.90     0.90     0.39
   7    5    20.02     1.10     0.47
   8    5    19.98     1.00     0.43
Grand Mean =     19.97, Std.Dev. =     0.359, Standard Error of Mean =
0.06
Mean Range =      0.89
Lower Control Limit =    0.000, Upper Control Limit =    1.876
```

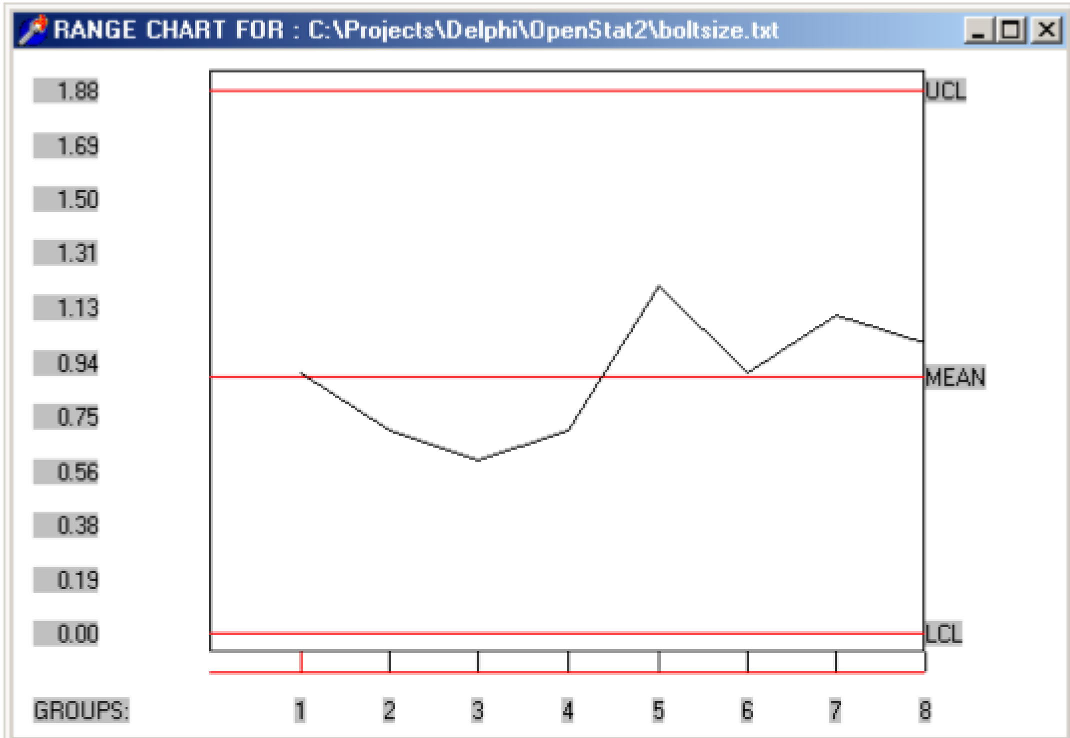


Fig. 12.5 Range Chart Plot

In the previous analysis using the XBAR chart procedure we found that the means of lots 3 and 6 were a meaningful distance from the target specification. In this chart we observed that lot 3 also had a larger range of values. The process appears out of control for lot 3 while for lot 6 it appears that the process was simply requiring adjustment toward the target value. In practice we would more likely see a pattern of increasing ranges as a machine becomes "loose" due to wear even though the averages may still be "on target".

S Control Chart

The sample standard deviation, like the range, is also an indicator of how much values vary in a sample. While the range reflects the difference between largest and smallest values in a sample, the standard deviation reflects the square root of the average squared distance around the mean of the values. We desire to reduce this variability in our processes so as to produce products as similar to one another as is possible. The S control chart plot the standard deviations of our sample lots and allows us to see the impact of adjustments and improvements in our manufacturing processes.

Examine the boltsize.txt data with the S Control Chart. Shown below is the specification form for the analysis and the results obtained:

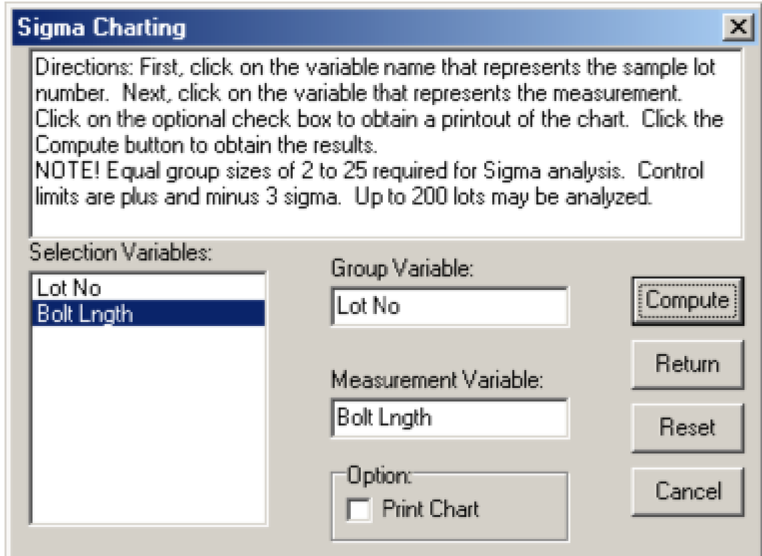


Fig. 12.6 Sigma Chart Dialog

```
Group Size Mean       Std.Dev.
_____ ____ _________ ________
   1    5    19.88      0.37
   2    5    19.90      0.29
   3    5    20.16      0.27
   4    5    20.08      0.29
   5    5    19.88      0.49
   6    5    19.90      0.39
   7    5    20.02      0.47
   8    5    19.98      0.43
Grand Mean =     19.97, Std.Dev. =     0.359, Standard Error of Mean =
0.06
Mean Sigma =      0.37
Lower Control Limit =     0.000, Upper Control Limit =     0.779
```

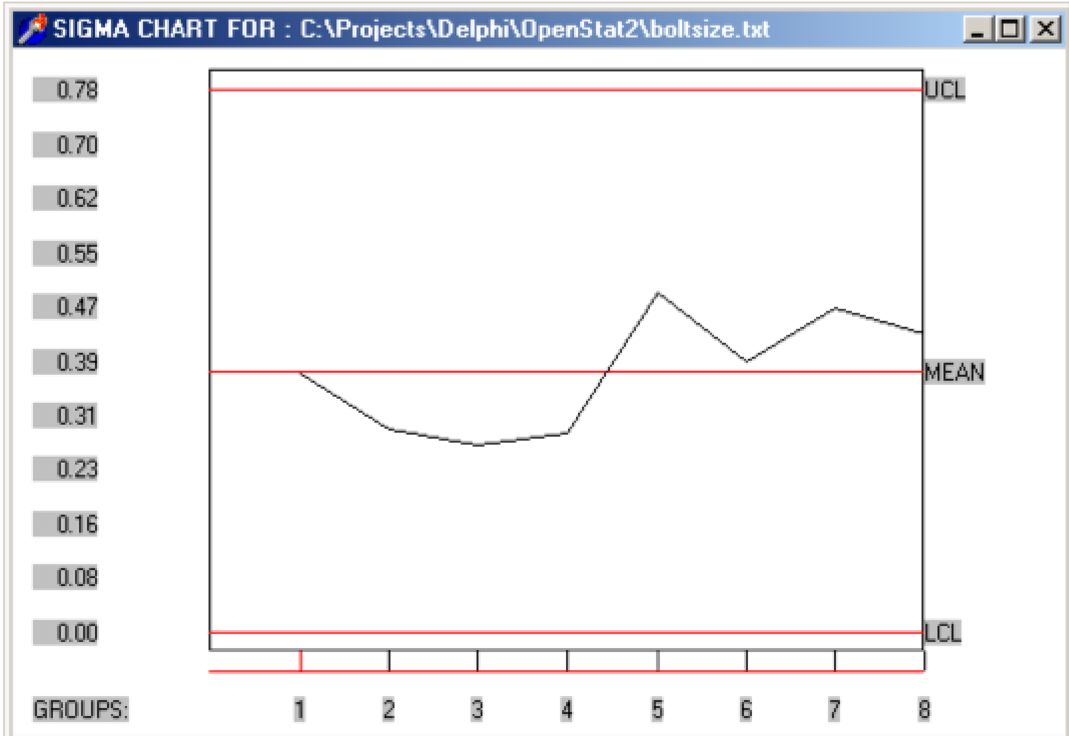


Fig. 12.7 Sigma Chart Plot

The pattern of standard deviations is similar to that of the Range Chart.

CUSUM Chart

The specification form for the CUSUM chart is shown below for the data file labeled boltsize.txt. We have specified our desire to detect shifts of 0.02 in the process and are using the 0.05 and 0.20 probabilities for the two types of errors.

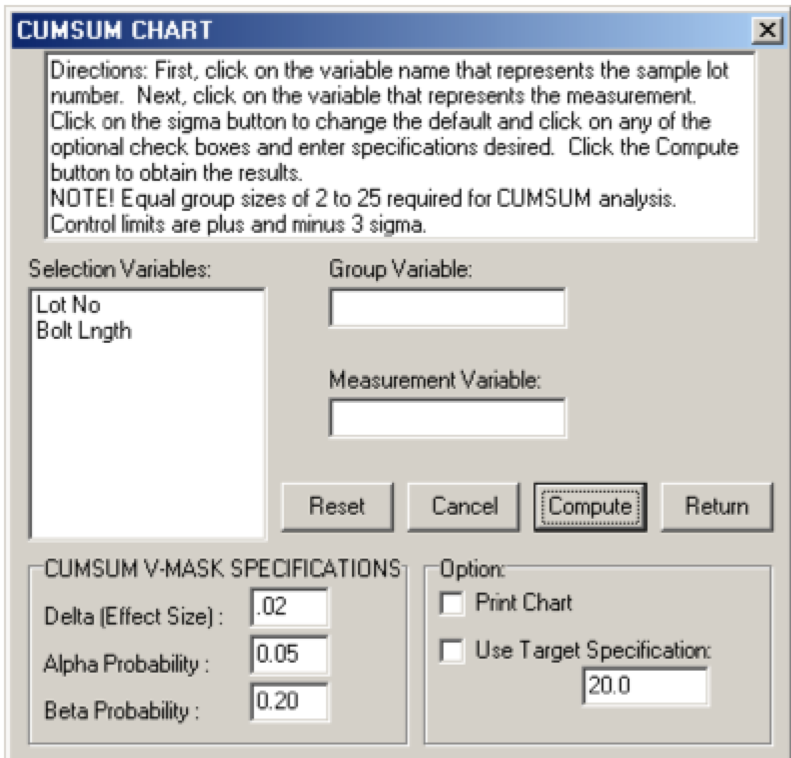


Fig. 12.8 CUMSUM Chart Dialog

```
CUMSUM Chart Results

Group Size Mean      Std.Dev.  Cum.Dev. of
                               mean from Target
_____ ____ ________  ________  ___________
    1   5    19.88     0.37        -0.10
    2   5    19.90     0.29        -0.18
    3   5    20.16     0.27         0.00
    4   5    20.08     0.29         0.10
    5   5    19.88     0.49         0.00
    6   5    19.90     0.39        -0.08
    7   5    20.02     0.47        -0.04
    8   5    19.98     0.43        -0.04
Mean of group deviations = -0.005
Mean of all observations = 19.975
Std. Dev. of Observations =    0.359
Standard Error of Mean =    0.057
Target Specification = 19.980
Lower Control Limit =   19.805, Upper Control Limit =   20.145
```

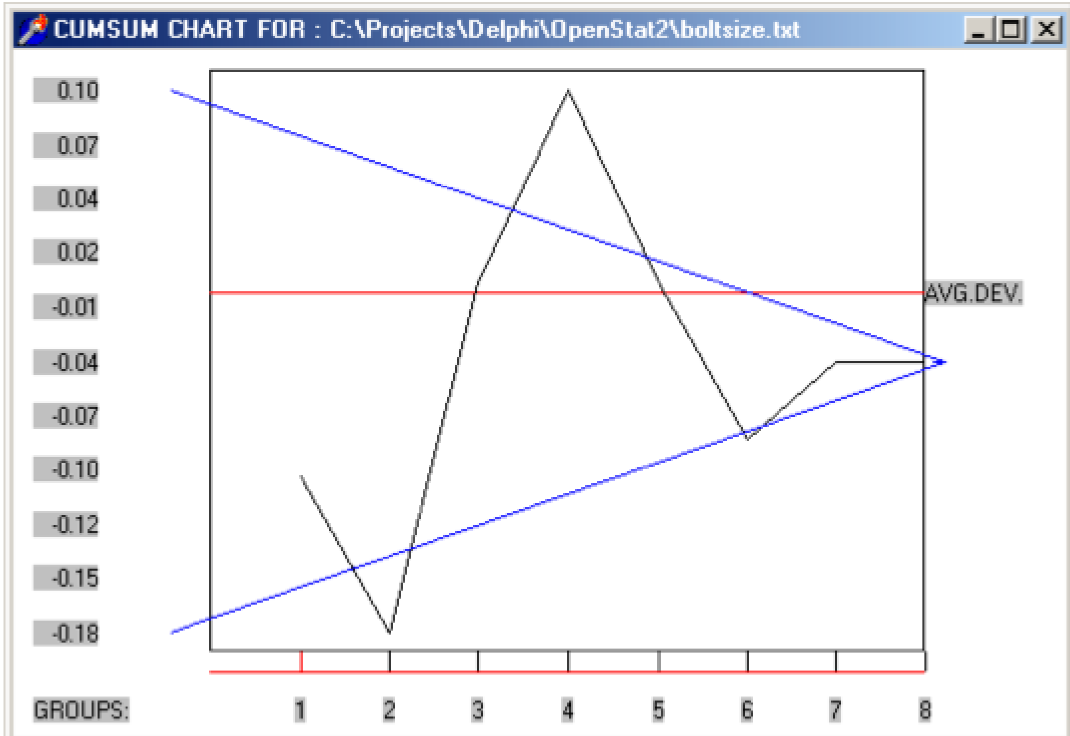


Fig. 12.9 CUMSUM Chart Plot

The results are NOT typical in that it appears that we have a process that is moving into control instead of out of control. Movement from lot 1 to 2 and from lot 3 to 4 indicate movement to out-of-control while the remaining values appear to be closer to in-control. If one checks the "Use the target value:" (of 20.0) the mask would indicate that lot 3 to 4 had moved to an out-of-control situation.

p Chart

To demonstrate the p Chart we will utilize a file labeled pchart.txt. Load the file and select the Analyses / Statistical Process Control / p Chart option. The specification form is shown below along with the results obtained after clicking the Compute Button:

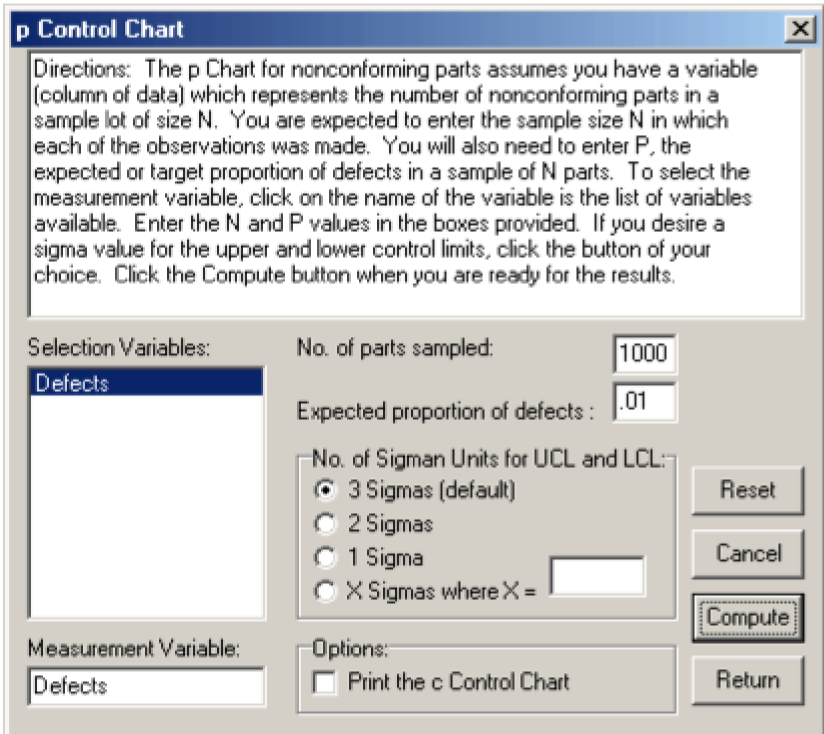


Fig. 12.10 p Control Chart Dialog

```
Target proportion = 0.0100
Sample size for each observation =   1000
Average proportion observed = 0.0116
Defects p Control Chart Results

Sample No.  Proportion
__________  __________
      1        0.012
      2        0.015
      3        0.008
      4        0.010
      5        0.004
      6        0.007
      7        0.016
      8        0.009
      9        0.014
     10        0.010
     11        0.005
     12        0.006
     13        0.017
     14        0.012
     15        0.022
     16        0.008
```

```
      17          0.010
      18          0.005
      19          0.013
      20          0.011
      21          0.020
      22          0.018
      23          0.024
      24          0.015
      25          0.009
      26          0.012
      27          0.007
      28          0.013
      29          0.009
      30          0.006
Target proportion = 0.0100
Sample size for each observation =    1000
Average proportion observed = 0.0116
```

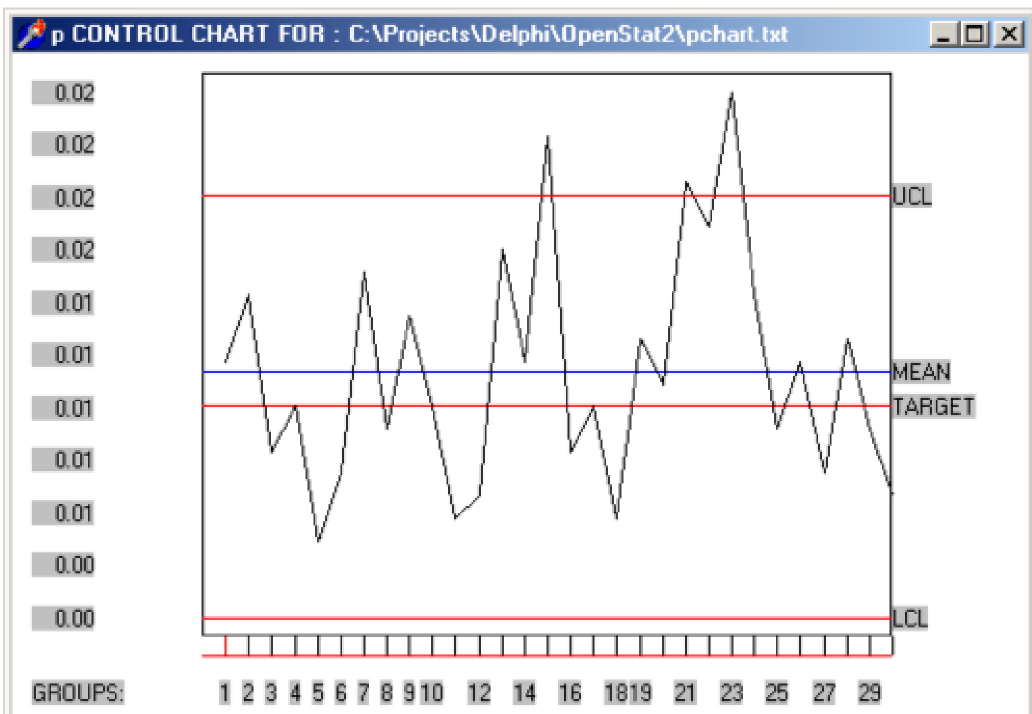


Fig. 12.11 p Control Chart Plot

Several of the sample lots (N = 1000) had disproportionately high defect rates and would bear further examination of what may have been occurring in the process at those points.

Defect (Non-conformity) c Chart

The previous section discusses the proportion of defects in samples (p Chart.) This section examines another defect process in which there is a count of defects in a sample lot. In this chart it is assumed that the occurrence of defects are independent, that is, the occurrence of a defect in one lot is unrelated to the occurrence in another lot. It is expected that the count of defects is quite small compared to the total number of parts potentially defective. For example, in the production of light bulbs, it is expected that in a sample of 1000 bulbs, only a few would be defective. The underlying assumed distribution model for the count chart is the Poisson distribution where the mean and variance of the counts are equal. Illustrated below is an example of processing a file labeled cChart.txt.

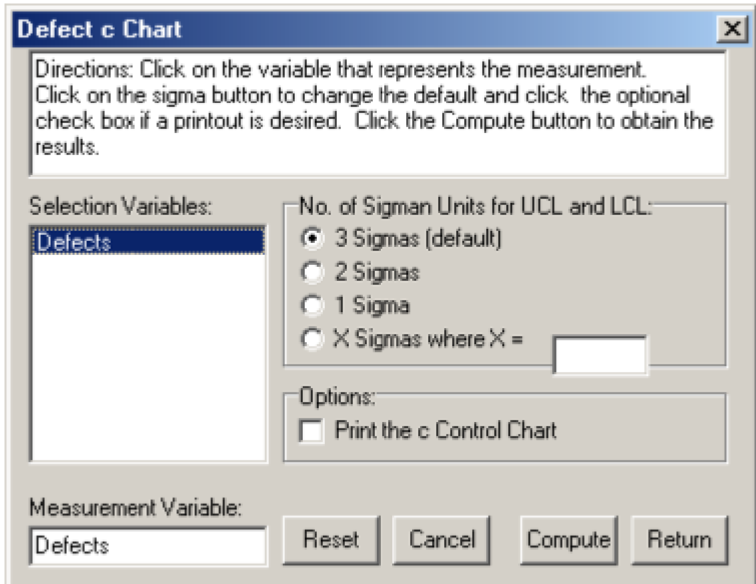


Fig. 12.12 Defect c Chart Dialog

```
Defects c Control Chart Results

Sample Number of
       Noncomformities

______ _______________
   1         7.00
   2         6.00
   3         6.00
   4         3.00
   5        22.00
   6         8.00
   7         6.00
   8         1.00
   9         0.00
  10         5.00
  11        14.00
  12         3.00
  13         1.00
  14         3.00
  15         2.00
  16         7.00
  17         5.00
  18         7.00
  19         2.00
  20         8.00
  21         0.00
  22         4.00
  23        14.00
  24         4.00
  25         3.00
Total Nonconformities =    141.00
```

```
No. of samples = 25
Poisson mean and variance =     5.640
Lower Control Limit =   -1.485, Upper Control Limit =   12.765
```

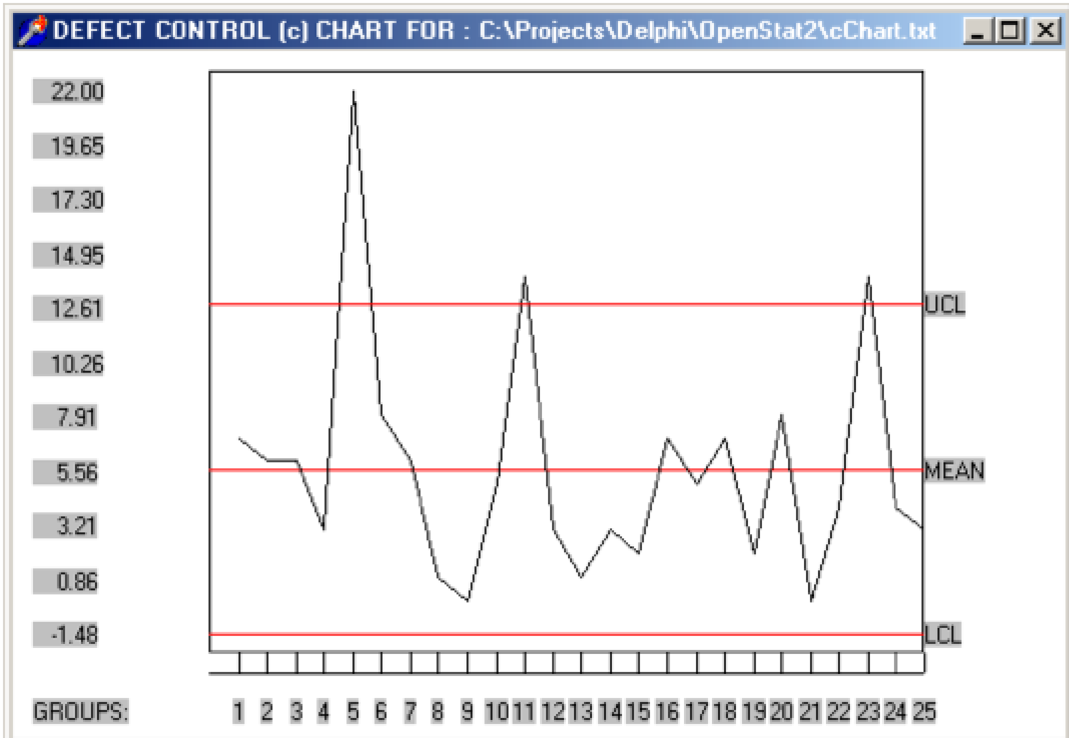


Fig. 12.13 Defect Control Chart Plot

The count of defects for three of the 25 objects is greater than the upper control limit of three standard deviations.

Defects Per Unit u Chart

The specification form and results for the computation following the click of the Compute button are shown below:

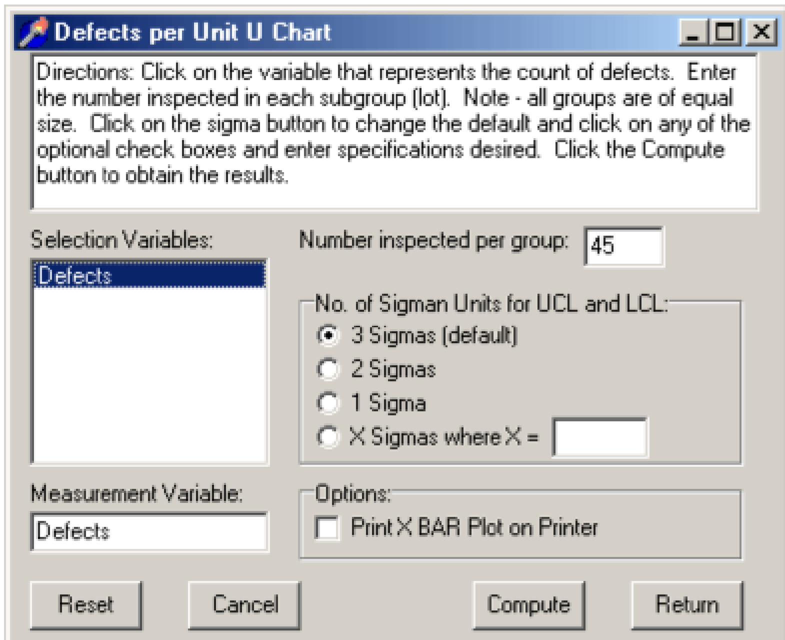


Fig. 12.14 Defects U Chart Dialog

Sample	No Defects	Defects Per Unit
1	36.00	0.80
2	48.00	1.07
3	45.00	1.00
4	68.00	1.51
5	77.00	1.71
6	56.00	1.24
7	58.00	1.29
8	67.00	1.49
9	38.00	0.84
10	74.00	1.64
11	69.00	1.53
12	54.00	1.20
13	56.00	1.24
14	52.00	1.16
15	42.00	0.93
16	47.00	1.04
17	64.00	1.42
18	61.00	1.36
19	66.00	1.47
20	37.00	0.82
21	59.00	1.31
22	38.00	0.84
23	41.00	0.91
24	68.00	1.51

```
  25         78.00        1.73
Total Nonconformities =  1399.00
No. of samples = 25
Def. / unit mean =    1.244 and variance =    0.166
Lower Control Limit =    0.745, Upper Control Limit =    1.742
```

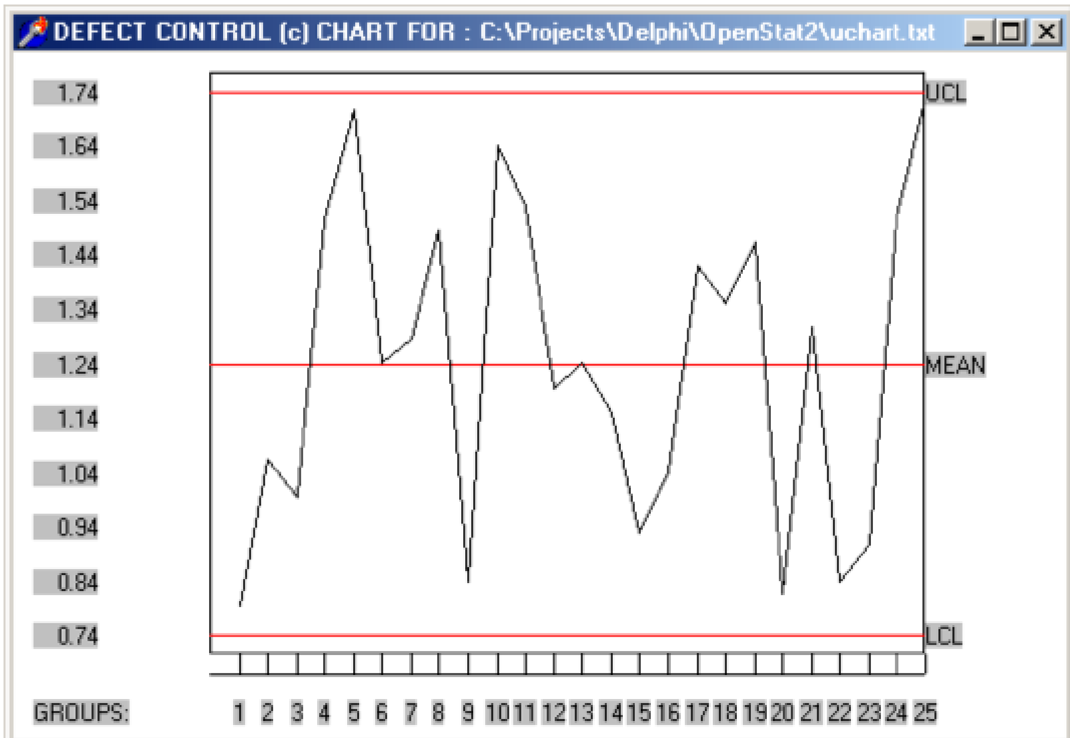


Fig. 12.15 Defect Control Chart Plot

In this example, the number of defects per unit are all within the upper and lower control limits.

Chapter 13. Financial

The Loan Amortization Procedure

To obtain a loan amortization schedule, click on this option to obtain the following dialog:

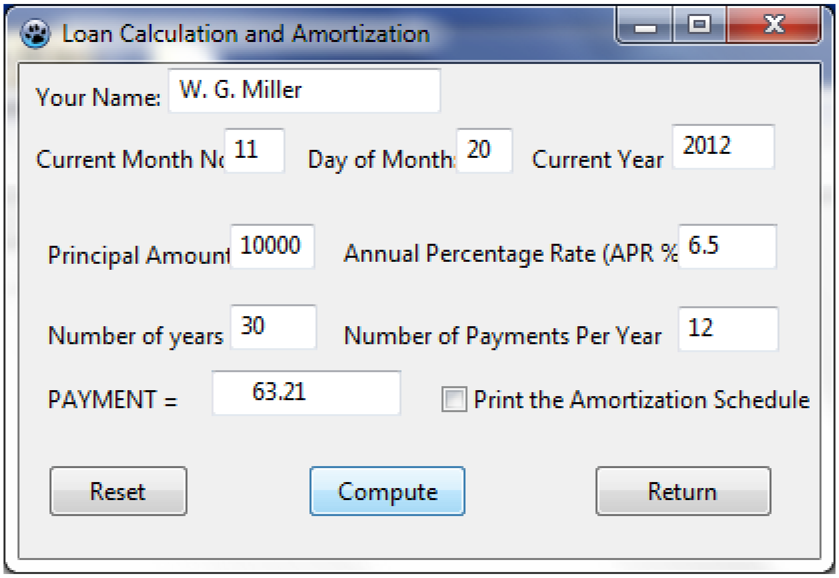


Fig. 13.1 The Loan Amortization Schedule Dialog

Sum of Years Depreciation

To obtain the sum of years depreciation, click on this option under the Financial menu and fill in the blanks of the dialog:

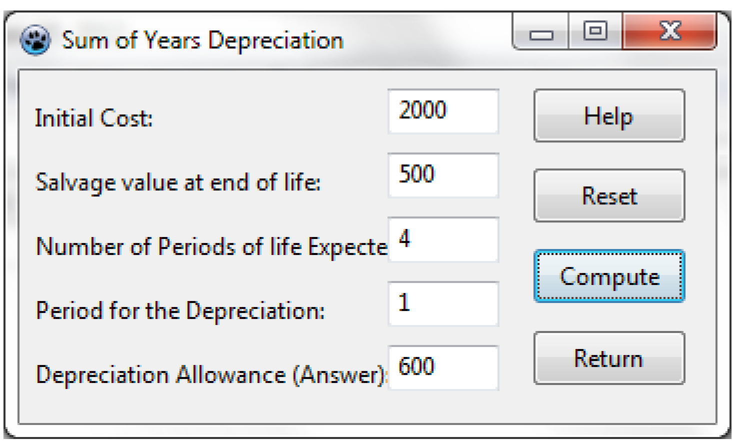


Fig. 13.2 The Sum of Years Depreciation dialog

Straight Line Depreciation

For a straight line depreciation, select the dialog from the Financial menu and complete the blanks before clicking the compute button:

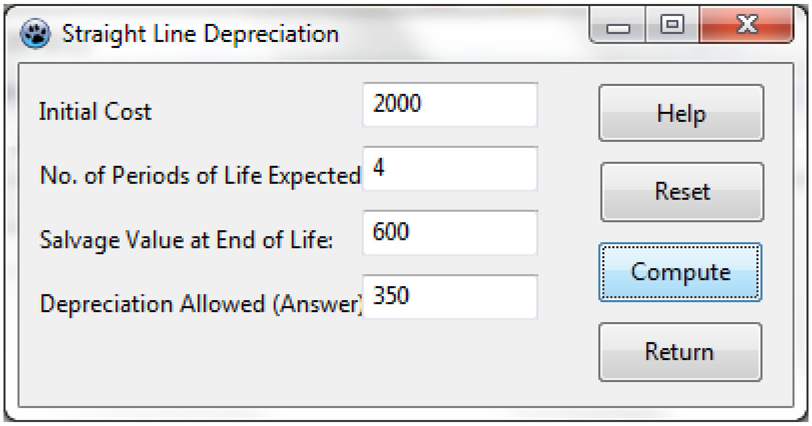


Fig. 13.3 The Straight Line Depreciation Dialog

Double Declining Value

You can obtain the double declining value by entering values on the dialog for this option under the financial menu:

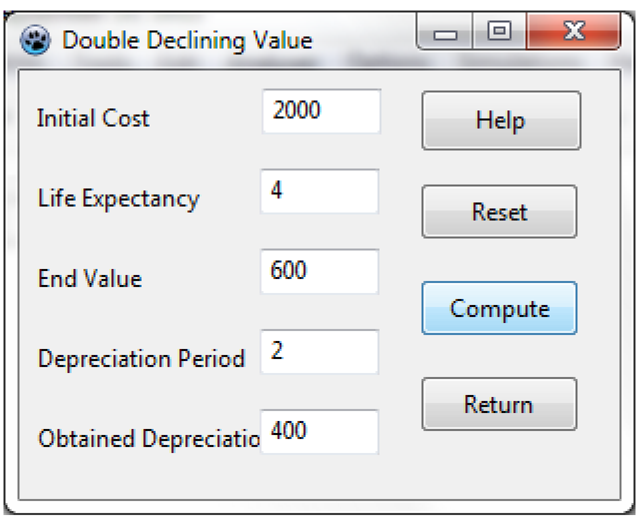


Fig. 13.4 The Double Declining Value Dialog

Chapter 14. Matrix Manipulation

Purpose of MatMan

MatMan was written to provide a platform for performing common matrix and vector operations. It is designed to be helpful for the student learning matrix algebra and statistics as well as the researcher needing a tool for matrix manipulation. If you are already a user of the OpenStat program, you can import files that you have saved with OpenStat into a grid of MatMan.

Using MatMan

When you first start the MatMan program, you will see the main program form below. This form displays four "grids" in which matrices, row or column vectors or scalars (single values) may be entered and saved. If a grid of data has already been saved, it can be retrieved into any one of the four grids. Once you have entered data into a grid, a number of operations can be performed depending on the type of data entered (matrix, vector or scalar.) Before performing an operation, you select the grid of data to analyze by clicking on the grid with the left mouse button. If the data in the selected grid is a matrix (file extension of .MAT) you can select any one of the matrix operations by clicking on the Matrix Operations "drop-down" menu at the top of the form. If the data is a row or column vector, select an operation option from the Vector Operations menu. If the data is a single value, select an operation from the Scalar Operations menu.

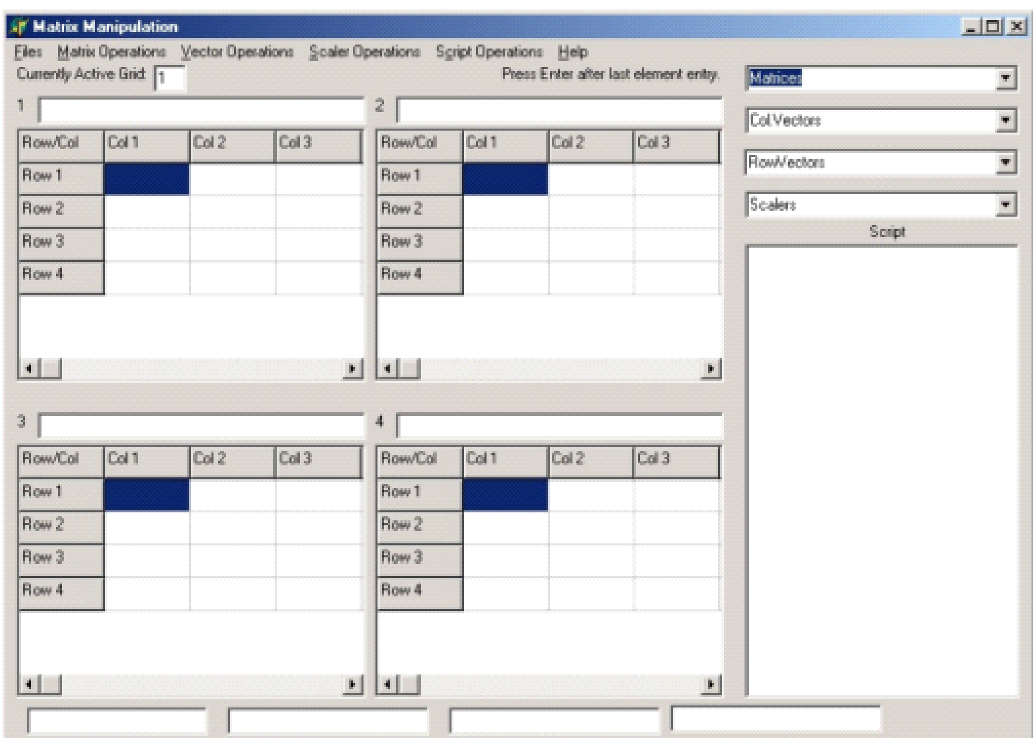


Fig. 14.1 The MatMan Dialog

Using the Combination Boxes

In the upper right portion of the MatMan main form, there are four "Combo Boxes". These boxes each contain a drop-down list of file names. The top box labeled "Matrix" contains the list of files containing matrices that have been created in the current disk directory and end with an extension of .MAT. The next two combo boxes contain similar lists of column or row vectors that have been created and are in the current disk directory. The last contains name of scalar files that have been saved in the current directory. These combo boxes provide documentation as to the names of current files already in use. In addition, they provide a "short-cut" method of opening a file and loading it into a selected grid.

Files Loaded at the Start of MatMan

Five types of files are loaded when you first start the execution of the MatMan program. The program will search for files in the current directory that have file extensions of .MAT, .CVE, .RVE, .SCA and .OPT. The first four types of files are simply identified and their names placed into the corresponding combination boxes of matrices, column vectors, row vectors and scalars. The last, options, is a file which contains only two integers: a 1 if the script should NOT contain File Open operations when it is generated or a 0 and a 1 if the script should NOT contain File Save operations when a script is generated or a 0. Since File Open and File Save operations are not actually executed when a script or script line is executed, they are in a script only for documentation purposes and may be left out.

Clicking the Matrix List Items

A list of Matrix files in the current directory will exist in the Matrix "Drop-Down" combination box when the MatMan program is first started. By clicking on one of these file names, you can directly load the referenced file into a grid of your selection.

Clicking the Vector List Items

A list of column and row vector files in the current directory will exist in the corresponding column vector or row vector "Drop-Down" combination boxes when the MatMan program is first started. By clicking a file name in one of these boxes, you can directly load the referenced file into a grid of your selection.

Clicking the Scalar List Items

When you click on the down arrow of the Scalar "drop-down" combination box, a list of file names appear which have been previously loaded by identifying all scalar files in the current directory. Also listed are any new scalar files that you have created during a session with MatMan. If you move your mouse cursor down to a file name and click on it, the file by that name will be loaded into the currently selected grid or a grid of your choice.

The Grids

The heart of all operations you perform involve values entered into the cells of a grid. These values may represent values in a matrix, a column vector, a row vector or a scalar. Each grid is like a spreadsheet. Typically, you select the first row and column cell by clicking on that cell with the left mouse key when the mouse cursor is positioned over that cell. To select a particular grid, click the left mouse button when the mouse cursor is positioned over any cell of that grid. You will then see that the grid number currently selected is displayed in a small text box in the upper left side of the form (directly below the menus.)

Operations and Operands

At the bottom of the form (under the grids) are four "text" boxes labeled Operation, Operand1, Operand2 and Operand3. Each time you perform an operation by use of one of the menu options, you will see an abbreviation of that operation in the Operation box. Typically there will be at least one or two operands related to that operation. The first operand is typically the name of the data file occupying the current grid and the second operand the name of the file containing the results of the operation. Some operations involve two grids, for example, adding two matrices. In these cases, the name of the two grid files involved will be in operands1 and operands2 boxes while the third operand box will contain the file for the results.

You will also notice that each operation or operand is prefixed by a number followed by a dash. In the case of the operation, this indicates the grid number from which the operation was begun. The numbers which prefix the operand labels indicate the grid in which the corresponding files were loaded or saved. The operation and operands are separated by a colon (:). When you execute a script line by double clicking an operation in the script list, the files are typically loaded into corresponding grid numbers and the operation performed.

Menus

The operations which may be performed on or with matrices, vectors and scalars are all listed as options under major menu headings shown across the top of the main form. For example, the File menu, when selected, provides a number of options for loading a grid with file data, saving a file of data from a grid, etc. Click on a menu heading and explore the options available before you begin to use MatMan. In nearly all cases, when you select a menu option you will be prompted to enter additional information. If you select an option by mistake you can normally cancel the operation.

Combo Boxes

Your main MatMan form contains what are known as "Drop-Down" combination boxes located on the right side of the form. There are four such boxes: The "Matrix" box, the "Column Vectors" box, the "Row Vectors" box and the "Scalars" box. At the right of each box is an arrow which, when clicked, results in a list of items "dropped-down" into view. Each item in a box represents the name of a matrix, vector or scalar file in the current directory or which has been created by one of the possible menu operations. By clicking on one of these items, you initiate the loading of the file containing the data for that matrix, vector or scalar. You will find this is a convenient alternative to use of the File menu for opening files which you have been working with. Incidentally, should you wish to delete an existing file, you may do so by selecting the "edit" option under the Script menu. The script editor lists all files in a directory and lets you delete a file by simply double-clicking the file name!

The Operations Script

Located on the right side of the main form is a rectangle which may contain operations and operands performed in using MatMan. This list of operations and their corresponding operands is known collectively as a "Script". If you were to perform a group of operations, for example, to complete a multiple regression analysis, you may want to save the script for reference or repeated analysis of another set of data. You can also edit the scripts that are created to remove operations you did not intend, change the file names referenced, etc. Scripts may also be printed.

Getting Help on a Topic

You obtain help on a topic by first selecting a menu item, grid or other area of the main form by placing the mouse over the item for which you want information. Once the area of interest is selected, press the F1 key on your keyboard. If a topic exists in the help file, it will be displayed. You can press the F1 key at any point to bring up the help file. A book is displayed which may be opened by double clicking it. You may also search for a topic using the help file index of keywords.

Scripts

Each time an operation is performed on grid data, an entry is made in a "Script" list shown in the right-hand portion of the form. The operation may have one to three "operands" listed with it. For example, the operation of finding the eigenvalues and eigenvectors of a matrix will have an operation of SVDInverse followed by the name of the matrix being inverted, the name of the eigenvalues matrix and the name of the eigenvectors matrix. Each part of the script entry is preceded by a grid number followed by a hyphen (-). A colon separates the parts of the entry (:). Once a series of operations have been performed the script that is produced can be saved. Saved scripts can be loaded at a later time and re-executed as a group or each entry executed one at a time. Scripts can also be edited and re-saved. Shown below is an example script for obtaining multiple regression coefficients.

```
CURRENT SCRIPT LISTING:

FileOpen:1-newcansas
1-ColAugment:newcansas:1-X
1-FileSave:1-X.MAT
1-MatTranspose:1-X:2-XT
2-FileSave:2-XT.MAT
2-PreMatxPostMat:2-XT:1-X:3-XTX
3-FileSave:3-XTX.MAT
3-SVDInverse:3-XTX.MAT:1-XTXINV
1-FileSave:1-XTXINV.MAT
FileOpen:1-XT.MAT
FileOpen:2-Y.CVE
1-PreMatxPostVec:1-XT.MAT:2-Y.CVE:3-XTY
3-FileSave:3-XTY.CVE
FileOpen:1-XTXINV.MAT
1-PreMatxPostVec:1-XTXINV.MAT:3-XTY:4-BETAS
4-FileSave:4-Bweights.CVE
```

Print

To print a script which appears in the Script List, move your mouse to the Script menu and click on the Print option. The list will be printed on the Output Form. At the bottom of the form is a print button that you can click with the mouse to get a hard-copy output.

Clear Script List

To clear an existing script from the script list, move the mouse to the Script menu and click the Clear option. Note: you may want to save the script before clearing it if it is a script you want to reference at a later time.

Edit the Script

Occasionally you may want to edit a script you have created or loaded. For example, you may see a number of Load File or Save File operations in a script. Since these are entered only for documentation and cannot actually be executed by clicking on them, they can be removed from the script. The result is a more compact and succinct script of operations performed. You may also want to change the name of files accessed for some operations or the name of files saved following an operation so that the same operations may be performed on a new set of data. To begin editing a script, move the mouse cursor to the Script menu and click on the Edit option. A new form appears which provides options for the editing. The list of operations appears on the left side of the form and an Options box appears in the upper right portion of the form. To edit a given script operation, click on the item to be edited and then click one of the option buttons. One option is to simply delete the item. Another is to edit (modify) the item. When that option is selected, the item is copied into an "Edit Box" which behaves like a miniature word processor. You can click on the text of an operation at any point in the edit box, delete characters following the cursor with the delete key, use the backspace key to remove characters in front of the cursor, and enter characters at the cursor. When editing is completed, press the return key to place the edited operation back into the script list from which it came.

Also on the Edit Form is a "Directory Box" and a "Files Box". Shown in the directory box is the current directory you are in. The files list box shows the current files in that directory. You can delete a file from any directory by simply double-clicking the name of the file in the file list. A box will pop up to verify that you want to delete the selected file. Click OK to delete the file or click Cancel if you do not want to delete the file. CAUTION! Be careful NOT to delete an important file like MATMAN.EXE, MATMAN.HLP or other system files (files with extensions of .exe, .dll, .hlp, .inf, etc.! Files which ARE safe to delete are those you have created with MatMan. These all end with an extension of .MAT, .CVE, .RVE ,.SCA or .SCR .

Load a Script

If you have saved a script of matrix operations, you can re-load the script for execution of the entire script of operations or execution of individual script items. To load a previously saved script, move the mouse to the Script menu and click on the Load option. Alternatively, you can go to the File menu and click on the Load Script option. Operation scripts are saved in a file as text which can also be read and edited with any word processing program capable of reading ASCII text files. For examples of scripts that perform statistical operations in matrix notation, see the help book entitled Script Examples.

Save a Script

Nearly every operation selected from one of the menus creates an entry into the script list. This script provides documentation of the steps performed in carrying out a sequence of matrix, vector or scalar operations. If you save the script in a file with a meaningful name related to the operations performed, that script may be "re-used" at a later time.

Executing a Script

You may quickly repeat the execution of a single operation previously performed and captured in the script. Simply click on the script item with the left mouse button when the cursor is positioned over the item to execute. Notice that you will be prompted for the name of the file or files to be opened and loaded for that operation. You can, of course, choose a different file name than the one or ones previously used in the script item. If you wish, you can also re-execute the entire script of operations. Move your mouse cursor to the Script menu and click on the Execute option. Each operation will be executed in sequence with prompts for file names appearing before execution each operation. Note: you will want to manually save the resulting file or files with appropriate names.

Script Options

File Open and File Save operations may or may not appear in a script list depending on options you have selected and saved. Since these two operations are *not* executed when a script is re-executed, it is not necessary that they be saved in a script (other than for documentation of the steps performed.) You can choose whether or not to have these operations appear in the script as you perform matrix, vector or scalar operations. Move your mouse cursor to the Script menu and click on the Options option. A pop-up form will appear on which you can elect to save or not save the File Open and File Save operations. The default (unchecked) option is to save these operations in a script. Clicking on an option tells the program to NOT write the operation to the script. Return to the MatMan main form by clicking the Return or Cancel button.

Files

When MatMan is first started it searches the current directory of your disk for any matrices, column vectors, row vectors or scalars which have previously been saved. The file names of each matrix, vector or scalar are entered into a drop-down list box corresponding to the type of data. These list boxes are located in the upper right portion of the main form. By first selecting one of the four grids with a click of the left mouse button and then clicking on one of the file names in a drop-down list, you can automatically load the file in the selected grid. Each time you save a grid of data with a new name, that file name is also added to the appropriate file list (Matrix, Column Vector, Row Vector or Scalar.)

At the top of the main form is a menu item labeled "Files". By clicking on the Files menu you will see a list of file options as shown in the picture below. In addition to saving or opening a file for a grid, you can also import an OpenStat .txt file, import a file with tab-separated values, import a file with comma separated values or import a file with spaces separating the values. All files saved with MatMan are ASCII text files and can be read (and edited if necessary) with any word processor program capable of reading ASCII files (for example the Windows Notepad program.)

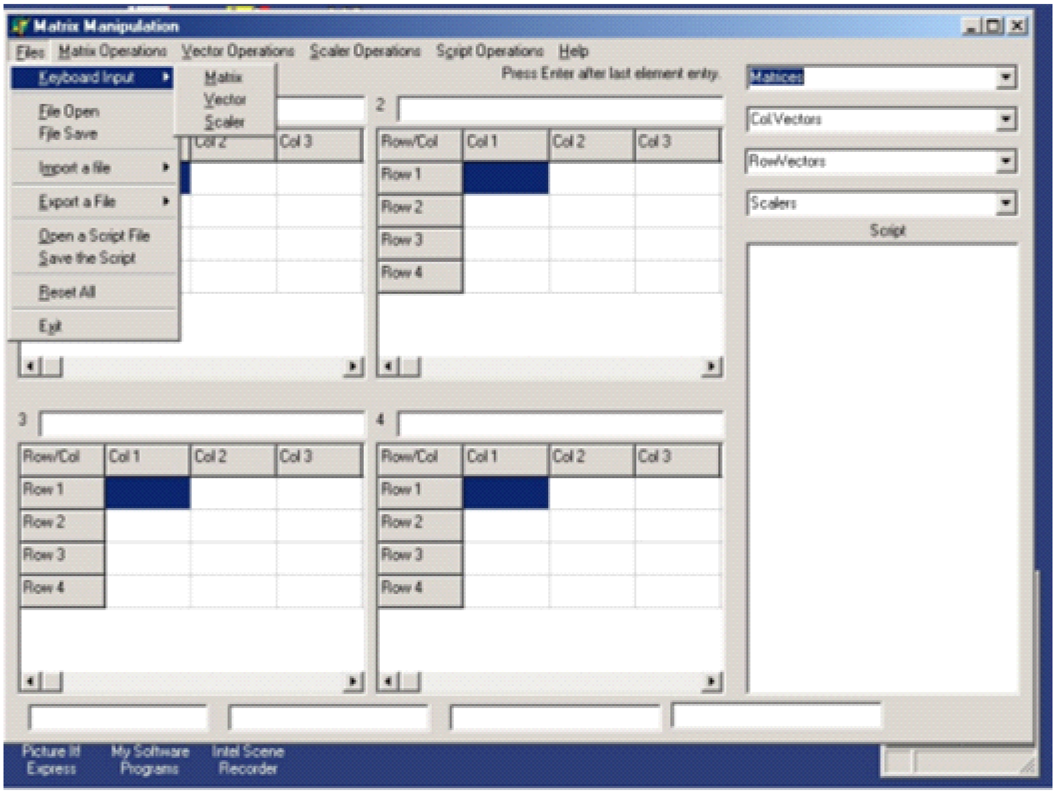


Fig. 14.2 Using the MatMan Files Menu

Keyboard Input

You can input data into a grid directly from the keyboard to create a file. The file may be a matrix, row vector, column vector or a scalar. Simply click on one of the four grids to receive your keystrokes. Note that the selected grid number will be displayed in a small box above and to the left of the grids. Next, click on the Files menu and move your cursor down to the Keyboard entry option. You will see that this option is expanded for you to indicate the type of data to be entered. Click on the type of data to be entered from the keyboard. If you selected a matrix, you will be prompted for the number of rows and columns of the matrix. For a vector, you will be prompted for the type (column or row) and the number of elements. Once the type of data to be entered and the number of elements are known, the program will "move" to the pre-selected grid and be waiting for your data entry. Click on the first cell (Row 1 and Column 1) and type your (first) value. Press the tab key to move to the next element in a row or, if at the end of a row, the first element in the next row. When you have entered the last value, instead of pressing the tab key, press the return key. You will be prompted to save the data. Of course, you can also go to the Files menu and click on the Save option. This second method is particularly useful if you are entering a very large data matrix and wish to complete it in several sessions.

File Open

If you have previously saved a matrix, vector or scalar file while executing the MatMan program, it will have been saved in the current directory (where the MatMan program resides.) MatMan saves data of a matrix type with a file extension of .MAT. Column vectors are saved with an extension of .CVE and row vectors saved with an extension of .RVE. Scalars have an extension of .SCA. When you click the File Open option in the File menu, a dialogue box appears. In the lower part of the box is an indication of the type of file. Click on this drop-down box to see the various extensions and click on the one appropriate to the type of file to be loaded. Once you have done that, the files listed in the files box will be only the files with that extension. Since the names of all matrix, vector and scalar files in the current directory are also loaded into the drop-down boxes in the upper right portion of the MatMan main form, you can also load a file by clicking on the name of the file in one of these boxes. Typically, you will be prompted for the grid number of the grid in which to load the file. The grid number is usually the one you have previously selected by clicking on a cell in one of the four grids.

File Save

Once you have entered data into a grid or have completed an operation producing a new output grid, you may save it by clicking on the save option of the File menu. Files are automatically saved with an extension which describes the type of file being saved, that is, with a .MAT, .CVE, .RVE or .SCA extension. Files are saved in the current directory unless you change to a different directory from the save dialogue box which appears when you are saving a file. It is recommended that you save files in the same directory (current directory) in which the MatMan program resides. The reason for doing this is that MatMan automatically loads the names of your files in the drop-down boxes for matrices, column vectors, row vectors and scalars.

Import a File

In addition to opening an existing MatMan file that has an extension of .MAT, .CVE, .RVE or .SCA, you may also *import* a file created by other programs. Many word processing and spread -sheet programs allow you to save a file with the data separated by tabs, commas or spaces. You can import any one of these types of files. Since the first row of data items may be the names of variables, you will be asked whether or not the first line of data contains variable labels.
You may also import files that you have saved with the OpenStat2 program. These files have an extension of .TXT or .txt when saved by the OpenStat2 program. While they are ASCII type text files, they contain a lot of information such as variable labels, long labels, format of data, etc. MatMan simply loads the variable labels, replacing the column labels currently in a grid and then loads numeric values into the grid cells of the grid you have selected to receive the data.

Export a File

You may wish to save your data in a form which can be imported into another program such as OpenStat, Excel, MicroSoft Word, WordPerfect, etc. Many programs permit you to import data where the data elements have been separated by a tab, comma or space character. The tab character format is particulary attractive because it creates an ASCII (American Standard Code for Information Interchange) file with clearly delineated spacing among values and which may be viewed by most word processing programs.

Open a Script File

Once you have performed a number of operations on your data you will notice that each operation has been "summarized" in a list of script items located in the script list on the right side of the MatMan form. This list of operations may be saved for later reference or re-execution in a file labeled appropriate to the series of operations. To re-open a script file, go to the File Menu and select the Open a Script File option. A dialogue box will appear. Select the type of file with an extension of .SCR and you will see the previously saved script files listed. Click on the one to load and press the OK button on the dialogue form. Note that if a script is already in the script list box, the new file will be added to the existing one. You may want to clear the script list box before loading a previously saved script. Clear the script list box by selecting the Clear option under the Script Operations menu.

Save the Script

Once a series of operations have been performed on your data, the operations performed will be listed in the Script box located to the right of the MatMan form. The series of operations may represent the completion of a data analysis such as multiple regression, factor analysis, etc. You may save this list of operations for future reference or re-execution. To save a script, select the Save Script option from the File Menu. A dialogue box will appear in which you enter the name of the file. Be sure that the type of file is selected as a .SCR file (types are selected in the drop-down box of the dialogue form.) A file extension of .SCR is automatically appended to the name you have entered. Click on the OK button to complete the saving of the script file.

Reset All

Occasionally you may want to clear all grids of data and clear all drop-down boxes of currently listed matrix, vector and scalar files. To do so, click the Clear All option under the Files Menu. Note that the script list box is NOT cleared by this operation. To clear a script, select the Clear operation under the Script Operations menu.

Entering Grid Data

Grids are used to enter matrices, vectors or scalars. Select a grid for data by moving the mouse cursor to the one of the grids and click the left mouse button. Move your mouse to the Files menu at the top of the form and click it with the left mouse button. Bring your mouse down to the Keyboard Input option. For entry of a matrix of values, click on the Matrix option. You will then be asked to verify the grid for entry. Press return if the grid number shown is correct or enter a new grid number and press return. You will then be asked to enter the name of your matrix (or vector or scalar.) Enter a descriptive name but keep it fairly short. A default extension of .MAT will automatically be appended to matrix files, a .CVE will be appended to column vectors, a .RVE appended to row vectors and a .SCA appended to a scalar. You will then be prompted for the number of rows and the number of columns for your data. Next, click on the first available cell labeled Col.1 and Row 1. Type the numeric value for the first number of your data. Press the tab key to move to the next column in a row (if you have more than one column) and enter the next value. Each time you press the tab key you will be ready to enter a value in the next cell of the grid. You can, of course, click on a particular cell to edit the value already entered or enter a new value. When you have entered the last data value, press the Enter key. A "Save" dialog box will appear with the name you previously chose. You can keep this name or enter a new name and click the OK button. If you later wish to edit values, load the saved file, make the changes desired and click on the Save option of the Files menu.

When a file is saved, an entry is made in the Script list indicating the action taken. If the file name is not already listed in one of the drop-down boxes (e.g. the matrix drop-down box), it will be added to that list.

Clearing a Grid

Individual grids are quickly reset to a blank grid with four rows and four columns by simply moving the mouse cursor over a cell of the grid and clicking the RIGHT mouse button. CAUTION! Be sure the data already in the grid has been saved if you do not want to lose it!

Inserting a Column

There may be occasions where you need to add another variable or column of data to an existing matrix of data. You may insert a new blank column in a grid by selecting the Insert Column operation under the Matrix Operations menu. First, click on an existing column in the matrix prior to or following the cell where you want the new column inserted. Click on the Insert Column option. You will be prompted to indicate whether the new column is to precede or follow the currently selected column. Indicate your choice and click the Return button.

Inserting a Row

There may be occasions where you need to add another subject or row of data to an existing matrix of data. You may insert a new blank row in a grid by selecting the Insert Row operation under the Matrix Operations menu. First, click on an existing row in the matrix prior to or following the cell where you want the new row inserted. Click on the Insert Row option. You will be prompted to indicate whether the new row is to precede or follow the number of the selected row. Indicate your choice and click the Return button.

Deleting a Column

To delete a column of data in an existing data matrix, click on the grid column to be deleted and click on the Delete Column option under the Matrix Operations menu. You will be prompted for the name of the new matrix to save. Enter the new matrix name (or use the current one if the previous one does not need to be saved) and click the OK button.

Deleting a Row

To delete a row of data in an existing data matrix, click on the grid row to be deleted and click on the Delete Row option under the Matrix Operations menu. You will be prompted for the name of the new matrix to save. Enter the new matrix name (or use the current one if the previous one does not need to be saved) and click the OK button.

Using the Tab Key

You can navigate through the cells of a grid by simply pressing the tab key. Of course, you may also click the mouse button on any cell to select that cell for data entry or editing. If you are at the end of a row of data and you press the tab key, you are moved to the first cell of the next row (if it exists.) To save a file press the Return key when located in the last row and column cell.

Using the Enter Key

If you press the Return key after entering the last data element in a matrix, vector or scalar, you will automatically be prompted to save the file. A "save" dialogue box will appear in which you enter the name of the file to save your data. Be sure the type of file to be saved is selected before you click the OK button.

Editing a Cell Value

Errors in data entry DO occur (after all, we are human aren't we?) You can edit a data element by simply clicking on the cell to be edited. If you double click the cell, it will be highlighted in blue at which time you can press the delete key to remove the cell value or enter a new value. If you simply wish to edit an existing value, click the cell so that it is NOT highlighted and move the mouse cursor to the position in the value at which you want to start editing. You can enter additional characters, press the backspace key to remove a character in front of the cursor or press the delete key to remove a character following the cursor. Press the tab key to move to the next cell or press the Return key to obtain the save dialogue box for saving your corrections.

Loading a File

Previously saved matrices, vectors or scalars are easily loaded into any one of the four grids. First select a grid to receive the data by clicking on one of the cells of the target grid. Next, click on the Open File option under the Files Menu. An "open" dialogue will appear which lists the files in your directory. The dialogue has a drop-down list of possible file types. Select the type for the file to be loaded. Only files of the selected type will then be listed. Click on the name of the file to load and click the OK button to load the file data.

Matrix Operations

Once a matrix of data has been entered into a grid you can elect to perform a number of matrix operations. The figure below illustrates the options under the Matrix Operations menu. Operations include:

Row Augment
Column Augment
Delete a Row
Delete a Column
Extract Col. Vector from Matrix
SVD Inverse
Tridiagonalize
Upper-Lower Decomposition
Diagonal to Vector
Determinant
Normalize Rows
Normalize Columns
Premultiply by : Row Vector; Matrix;Scaler
Postmultiply by : Column Vector; Matrix
Eigenvalues and Vectors
Transpose
Trace
Matrix A + Matrix B
Matrix A - Matrix B
Print

Printing

You may elect to print a matrix, vector, scalar or file. When you do, the output is placed on an "Output" form. At the bottom of this form is a button labeled "Print" which, if clicked, will send the contents of the output form to the printer. Before printing this form, you may type in additional information, edit lines, cut and paste lines and in general edit the output to your liking. Edit operations are provided as icons at the top of the form. Note that you can also save the output to a disk file, load another output file and, in general, use the output form as a word processor.

Row Augment

You may add a row of 1's to a matrix with this operation. When the transpose of such an augmented matrix is multiplied times this matrix, a cell will be created in the resulting matrix, which contains the number of columns in the augmented matrix.

Column Augmentation

You may add a column of 1's to a matrix with this operation. When the transpose of such an augmented matrix is multiplied times this matrix, a cell will be created in the resulting matrix, which contains the number of rows in the augmented matrix. The procedure for completing a multiple regression analysis often involves column augmentation of a data matrix containing a row for each object (e.g. person) and column cells containing independent variable values. The column of 1's created from the Column Augmentation process ends up providing the intercept (regression constant) for the analysis.

Extract Col. Vector from Matrix

In many statistics programs the data matrix you begin with contains columns of data representing independent variables and one or more columns representing dependent variables. For example, in multiple regression analysis, one column of data represents the dependent variable (variable to be predicted) while one or more columns represent independent variables (predictor variables.) To analyze this data with the MatMan program, one would extract the dependent variable and save it as a column vector for subsequent operations (see the sample multiple regression script.) To extract a column vector from a matrix you first load the matrix into one of the four grids, click on a cell in the column to be extracted and then click on the Extract Col. Vector option under the Matrix Operations menu.

SVDInverse

A commonly used matrix operation is the process of finding the inverse (reciprocal) of a symmetric matrix. A variety of methods exist for obtaining the inverse (if one exists.) A common problem with some inverse methods is that they will not provide a solution if one of the variables is dependent (or some combination of) on other variables (rows or columns) of the matrix. One advantage of the "Singular Value Decomposition" method is that it typically provides a solution even when one or more dependent variables exist in the matrix. The offending variable(s) are essentially replaced by zeroes in the row and column of the dependent variable. The resulting inverse will NOT be the desired inverse.

To obtain the SVD inverse of a matrix, load the matrix into a grid and click on the SVDInverse option from the Matrix Operations menu. The results will be displayed in grid 1 of the main form. In addition, grids 2 through 4 will contain additional information which may be helpful in the analysis. Figures 1 and 2 below illustrate the results of inverting a 4 by 4 matrix, the last column of which contains values that are the sum of the first three column cells in each row (a dependent variable.)

When you obtain the inverse of a matrix, you may want to verify that the resulting inverse is, in fact, the reciprocal of the original matrix. You can do this by multiplying the original matrix times the inverse. The result should be a matrix with 1's in the diagonal and 0's elsewhere (the identity matrix.) Figure 3 demonstrates that the inverse was NOT correct, that is, did not produce an identity matrix when multiplied times the original matrix.

```
Figure 1. DepMat.MAT From Grid Number 1


                     Columns
                   Col.1        Col.2        Col.3        Col.4
Rows
```

1	5.000	11.000	2.000	18.000
2	11.000	2.000	4.000	17.000
3	2.000	4.000	1.000	7.000
4	18.000	17.000	7.000	1.000

Figure 2. DepMatInv.MAT From Grid Number 1

Rows	Columns Col.1	Col.2	Col.3	Col.4
1	0.584	0.106	-1.764	0.024
2	0.106	-0.068	-0.111	0.024
3	-1.764	-0.111	4.802	0.024
4	0.024	0.024	0.024	-0.024

Figure 3. DepMatxDepMatInv.MAT From Grid Number 3

Rows	Columns Col.1	Col.2	Col.3	Col.4
1	1.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000
3	0.000	0.000	1.000	0.000
4	1.000	1.000	1.000	0.000

NOTE! This is NOT an Identity matrix.

Tridiagonalize

In obtaining the roots and vectors of a matrix, one step in the process is frequently to reduce a symetric matrix to a tri-diagonal form. The resulting matrix is then solved more readily for the eigenvalues and eigenvectors of the original matrix. To reduce a matrix to its tridiagonal form, load the original matrix in one of the grids and click on the Tridiagonalize option under the Matrix Operations menu.

Upper-Lower Decomposition

A matrix may be decomposed into two matrices: a lower matrix (one with zeroes above the diagonal) and an upper matrix (one with zeroes below the diagonal matrix.) This process is sometimes used in obtaining the inverse of a matrix. The matrix is first decomposed into lower and upper parts and the columns of the inverse solved one at a time using a routine that solves the linear equation A X = B where A is the upper/lower decomposition matrix, B are known result values of the equation and X is solved by the routine. To obtain the LU decomposition, enter or load a matrix into a grid and select the Upper-Lower Decomposition option from the Matrix Operations menu.

Diagonal to Vector

In some matrix algebra problems it is necessary to perform operations on a vector extracted from the diagonal of a matrix. The Diagonal to Vector operation extracts the the diagonal elements of a matrix and creates a new column vector with those values. Enter or load a matrix into a grid and click on the Diagonal to Vector option under the Matrix Operations menu to perform this operation.

Determinant

The determinant of a matrix is a single value characterizing the matrix values. A singular matrix (one for which the inverse does not exist) will have a determinant of zero. Some ill-conditioned matrices will have a determinant close to zero. To obtain the determinant of a matrix, load or enter a matrix into a grid and select the Determinant option from among the Matrix Operations options. Shown below is the determinant of a singular matrix (row/column 4 dependent on columns 1 through 3.)

```
                     Columns
                   Col.1         Col.2         Col.3         Col.4
Rows
    1              5.000        11.000         2.000        18.000
    2             11.000         2.000         4.000        17.000
    3              2.000         4.000         1.000         7.000
    4             18.000        17.000         7.000        42.000
```

```
                 Columns
               Col 1
Rows
    1          0.000
```

Normalize Rows or Columns

In matrix algebra the columns or rows of a matrix often represent vectors in a multi-dimension space. To make the results more interpretable, the vectors are frequently scaled so that the vector length is 1.0 in this "hyper-space" of k-dimensions. This scaling is common for statistical procedures such as Factor Analysis, Principal Component Analysis, Discriminant Analysis, Multivariate Analysis of Variance, etc. To normalize the row (or column) vectors of a matrix such as eigenvalues, load the matrix into a grid and select the Normalize Rows (or Normalize Columns) option from the Matrix Operations menu.

Pre-Multiply by:

A matrix may be multiplied by a row vector, another matrix or a single value (scalar.) When a row vector with N columns is multiplied times a matrix with N rows, the result is a row vector of N elements. When a matrix of N rows and M columns is multiplied times a matrix with M rows and Q columns, the result is a matrix of N rows and Q columns. Multiplying a matrix by a scalar results in each element of the matrix being multiplied by the value of the scalar.

To perform the pre-multiplication operation, first load two grids with the values of a matrix and a vector, matrix or scaler. Click on a cell of the grid containing the matrix to insure that the matrix grid is selected. Next, select the Pre-Multipy by: option and then the type of value for the pre-multiplier in the sub-options of the Matrix Operations menu. A dialog box will open asking you to enter the grid number of the matrix to be multiplied. The default value is the selected matrix grid. When you press the OK button another dialog box will prompt you for the grid number containing the row vector, matrix or scalar to be multiplied times the matrix. Enter the grid number for the pre-multiplier and press return. Finally, you will be prompted to enter the grid number where the results are to be displayed. Enter a number different than the first two grid numbers entered. You will then be prompted for the name of the file for saving the results.

Post-Multiply by:

A matrix may be multiplied times a column vector or another matrix. When a matrix with N rows and Q columns is multiplied times a column vector with Q rows, the result is a column vector of N elements. When a matrix of N rows and M columns is multiplied times a matrix with M rows and Q columns, the result is a matrix of N rows and Q columns.

To perform the post-multiplication operation, first load two grids with the values of a matrix and a vector or matrix. Click on a cell of the grid containing the matrix to insure that the matrix grid is selected. Next, select the Post-Multiply by: option and then the type of value for the post-multiplier in the sub-options of the Matrix Operations menu. A dialog box will open asking you to enter the grid number of the matrix multiplier. The default value is the selected matrix grid. When you press the OK button another dialog box will prompt you for the grid number containing the column vector or matrix. Enter the grid number for the post-multiplier and press return. Finally, you will be prompted to enter the grid number where the results are to be displayed. Enter a number different than the first two grid numbers entered. You will then be prompted for the name of the file for saving the results.

Eigenvalues and Vectors

Eigenvalues represent the k roots of a polynomial constructed from k equations. The equations are represented by values in the rows of a matrix. A typical equation written in matrix notation might be:

$$Y = B X$$

where X is a matrix of known "independent" values, Y is a column vector of "dependent" values and B is a column vector of coefficients which satisfies specified properties for the solution. An example is given when we solve for "least-squares" regression coefficients in a multiple regression analysis. In this case, the X matrix contains cross-products of k independent variable values for N cases, Y contains known values obtained as the product of the transpose of the X matrix times the N values for subjects and B are the resulting regression coefficients.

In other cases we might wish to transform our matrix X into another matrix V which has the property that each column vector is "orthogonal" to (un-correlated) with the other column vectors. For example, in Principal Components analysis, we seek coefficients of vectors that represent new variables that are uncorrelated but which retain the variance represented by variables in the original matrix. In this case we are solving the equation

$$VXV^T = \lambda$$

X is a symmetric matrix and λ are roots of the matrix stored as diagonal values of a matrix. If the columns of V are normalized then $V V^T = I$, the identity matrix.

Transpose

The transpose of a matrix or vector is simply the creation of a new matrix or vector where the number of rows is equal to the number of columns and the number of columns equals the number of rows of the original matrix or vector. For example, the transpose of the row vector [1 2 3 4] is the column vector:

1
2
3
4

Similarly, given the matrix of values:

1	2	3
4	5	6

the transpose is:

1	4
2	5
3	6

You can transpose a matrix by selecting the grid in which your matrix is stored and clicking on the Transpose option under the Matrix Operations menu. A similar option is available under the Vector Operations menu for vectors.

Trace

The trace of a matrix is the sum of the diagonal values.

Matrix A + Matrix B

When two matrices of the same size are added, the elements (cell values) of the first are added to corresponding cells of the second matrix and the result stored in a corresponding cell of the results matrix. To add two matrices, first be sure both are stored in grids on the main form. Select one of the grid containing a matrix and click on the Matrix A + Matrix B option in the Matrix Operations menu. You will be prompted for the grid numbers of each matrix to be added as well as the grid number of the results. Finally, you will be asked the name of the file in which to save the results.

Matrix A - Matrix B

When two matrices of the same size are subtracted, the elements (cell values) of the second are subtracted from corresponding cells of the first matrix and the result stored in a corresponding cell of the results matrix. To subtract two matrices, first be sure both are stored in grids on the main form. Select one of the grids containing the matrix from which another will be subtracted and click on the Matrix A - Matrix B option in the Matrix Operations menu. You will be prompted for the grid numbers of each matrix as well as the grid number of the results. Finally, you will be asked the name of the file in which to save the results.

Print

To print a matrix be sure the matrix is loaded in a grid, the grid selected and then click on the print option in the Matrix Operations menu. The data of the matrix will be shown on the output form. To print the output form on your printer, click the Print button located at the bottom of the output form.

Vector Operations

A number of vector operations may be performed on both row and column vectors. Shown below is the main form with the Vector Operations menu selected. The operations you may perform are:

Transpose
Multiply by Scalar
Square Root of Elements
Reciprocal of Elements
Print
Row Vec. x Col. Vec.
Col. Vec x Row Vec.

Vector Transpose

The transpose of a matrix or vector is simply the interchange of rows with columns. Transposing a matrix results in a matrix with the first row being the previous first column, the second row being the previous second column, etc. A column vector becomes a row vector and a row vector becomes a column vector. To transpose a vector, click on the grid where the vector resides that is to be transposed. Select the Transpose Option from the Vector Operations menu and click it. Save the transposed vector in a file when the save dialogue box appears.

Multiply a Vector by a Scalar

When you multiply a vector by a scalar, each element of the vector is multiplied by the value of that scalar. The scalar should be loaded into one of the grids and the vector in another grid. Click on the

Multiply by a Scalar option under the Vector Operations menu. You will be prompted for the grid numbers containing the scalar and vector. Enter those values as prompted and click the return button following each. You will then be presented a save dialogue in which you enter the name of the new vector.

Square Root of Vector Elements

You can obtain the square root of each element of a vector. Simply select the grid with the vector and click the Square Root option under the Vector Operations menu. A save dialogue will appear after the execution of the square root operations in which you indicate the name of your new vector. Note - you cannot take the square root of a vector that contains a negative value - an error will occur if you try.

Reciprocal of Vector Elements

Several statistical analysis procedures involve obtaining the reciprocal of the elements in a vector (often the diagonal of a matrix.) To obtain reciprocals, click on the grid containing the vector then click on the Reciprocal option of the Vector Operations menu. Of course, if one of the elements is zero, an error will occur! If valid values exist for all elements, you will then be presented a save dialogue box in which you enter the name of your new vector.

Print a Vector

Printing a vector is the same as printing a matrix, scalar or script. Simply select the grid to be printed and click on the Print option under the Vector Operations menu. The printed output is displayed on an output form. The output form may be printed by clicking the print button located at the bottom of the form.

Row Vector Times a Column Vector

Multiplication of a column vector by a row vector will result in a single value (scalar.) Each element of the row vector is multiplied times the corresponding element of the column vector and the products are added. The number of elements in the row vector must be equal to the number of elements in the column vector. This operation is sometimes called the "dot product" of two vectors. Following execution of this vector operation, you will be shown the save dialogue for saving the resulting scalar in a file.

Column Vector Times Row Vector

When you multiply a column vector of k elements times a row vector of k elements, the result is a k by k matrix. In the resulting matrix each row by column cell is the product of the corresponding column element of the row vector and the corresponding row element of the column vector. The result is equivalent to multiplying a k by 1 matrix times a 1 by k matrix.

Scalar Operations

The operations available in the Scalar Operations menu are:

Square Root
Reciprocal
Scalar x Scalar
Print

Square Root of a Scalar

Selecting this option under the Scalar Operations menu results in a new scalar that is the square root of the original scalar. The new value should probably be saved in a different file than the original scalar. Note that you will get an error message if you attempt to take the square root of a negative value.

Reciprocal of a Scalar

You obtain the reciprocal of a scalar by selecting the Reciprocal option under the Scalar Operations menu. You will obtain an error if you attempt to obtain the reciprocal of a value zero. Save the new scalar in a file with an appropriate label.

Scalar Times a Scalar

Sometimes you need to multiply a scalar by another scalar value. If you select this option from the Scalar Operations menu, you will be prompted for the value of the muliplier. Once the operation has been completed you should save the new scalar product in a file appropriately labeled.

Print a Scalar

Select this option to print a scalar residing in one of the four grids that you have selected. Notice that the output form contains all objects that have been printed. Should you need to print only one grid's data (matrix, vector or scalar) use the Clear All option under the Files menu.

BIBLIOGRAPHY

.

1. Afifi, A. A. and Azen, S. P. *Statistical Analysis. A Computer Oriented Approach*. New York: Academic Press, Inc. 1972.

2. Anderberg, Michael R. *Cluster Analysis for Applications*. New York: Academic Press, 1973.

3. Bennett, Spencer and Bowers, David. *An Introduction to Multivariate Techniques for Social and Behavioral Sciences*. New York: John Wiley and Sons, Inc., 1977.

4. Besterfield, Dale H. *Quality Control*, *2nd Ed*. Englewood Ciffs, N.J.: Prentice-Hall, Inc., 1986.

5. Bishop, Yvonne M., Fienberg, Stephen E. and Holland, Paul W. *Discrete Multivariate Analysis. Theory and Practice*. Cambridge, Mass.: The MIT Press 1975.

6. Blommers, Paul J. and Forsyth, Robert A. *Elementary Statistical Methods in Psychology and Education, 2nd Ed*. Boston, Mass.: Houghton Mifflin Company, 1977.

7. Borg, Walter R. and Gall, Meridith Damien. *Educational Research. An Introduction, 5th Ed*. New York: Longman, Inc., 1989.

8. Brockwell, Peter J. and Davis, Richard A. *Introduction to Time Series and Forecasting*. New York: Springer-Verlag New York Inc., 1996.

9. Bruning, James L. and Kintz, B. L. *Computational Handbook of Statistics, 2nd Ed*. Glenview, Ill.: Scott, Foresman and Company, 1977.

10. Campbell, Donald T. and Stanley, Julian C. *Experimental and Quasi-Experimental Designs for Research*. Chicago, Ill.: Rand McNally College Publishing Company, 1963.

11. Chapman, Douglas G. and Schaufele, Ronald A. *Elementary Probability Models and Statistical Inference*. Waltham, Mass.: Ginn-Blaisdell, 1970.

12. Cody, Ronald P. and Smith, Jeffrey K. *Applied Statistics and the SAS Programming Language, 4th Ed*. Upper Saddle River, N.J.: Prentice Hall, 1997.

13. Cohen, Jacob and Cohen, Patricia. *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

14. Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences, 2nd Ed*., Hillsdale, N.J.: Lawrence Erlbaum Associates, 1988.

15. Comrey, Andrew L. *A First Course in Factor Analysis*. New York: Academic Press, Inc., 1973.

16. Cook, Thomas D. and Campbell, Donald T. *Quasi-Experimentation. Design and Analysis Issues for Field Settings*. Chicago, Ill.: Rand McNally College Publishing Company, 1979.

17. Cooley, William W. and Lohnes, Paul R. *Multivariate Data Analysis*. New York: John Wiley and Sons, Inc., 1971.

18. Crocker, Linda and Algina, James. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston, 1986.

19. Diekhoff, George M. *Basic Statistics for the Social and Behavioral Sciences*. Upper Sadle River, N.J., Prentice Hall, Inc. 1996.

20. Edwards, Allen L. *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts, Inc., 1957.

21. Efromovich, Sam. *Nonparametric Curve Estimation. Methods, Theory, and Applications*. New York: Springer-Verlag, 1999.

22. Ferrguson, George A. *Statistical Analysis in Psychology and Education, 2nd Ed*. New York: McGraw-Hill Book Company, 1966.

23. Fienberg, Stephen E. *The Analysis of Cross-Classified Categorical Data, 2nd Ed*. Cambridge, Mass.: The MIT Press, 1980.

24. Fox, John. *Multiple and Generalized Nonparametric Regression*. Thousand Oaks, Cal.: Sage Publications, Inc., 2000.

25. Freund, John E. and Walpole, Ronald E. *Mathematical Statistics, 4th Ed*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1987.

26. Fruchter, Benjamin. Introduction to Factor Analysis. Princeton, N.J.: D. Van Nostrand Company, Inc., 1954.

27. Gay, L. R. *Educational Research. Competencies for Analysis and Application, 4th Ed*. New York: Macmillan Publishing Company, 1992.

28. Glass, Gene V. and Stanley, Julian C. *Statistical Methods in Education and Psychology*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.

29. Gottman, John M. and Leiblum, Sandra R. *How to do Psychotherapy and How to Evaluate It. A Manual for Beginners*. New York: Holt, Rinehart and Winston, Inc., 1974.

30. Guertin, Wilson H. and Bailey, Jr., John P. *Introduction to Modern Factor Analysis*. Ann Arbor, Mich.: Edwards Brothers, Inc., 1970.

31. Gulliksen, Harold. *Theory of Mental Tests*. New York: John Wiley and Sons, Inc., 1950.

32. Hambleton, Ronald K. and Swaminathan, Hariharan. *Item Response Theory. Principles and Applications*. Boston, Mass.: Kluwer-Nijhoff Publishing, 1985.

33. Hansen, Bertrand L. and Chare, Prabhakar M. *Quality Control and Applications*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1987.

34. Harman, Harry H. *Modern Factor Analysis*. Chicago, Ill.: The University of Chicago Press, 1960.

35. Hays, William L. *Statistics for Psychologists*. New York: Holt, Rinehart and Winston, Inc., 1963.

36. Heise, David R. *Causal Analysis*. New York: John Wiley and Sons, Inc., 1975.

37. Hinkle, Dennis E., Wiersma, William and Jurs, Stephen G. *Applied Statistics for the Behavioral Sciences*, *2nd Edition*. Boston, Mass.: Houghton Mifflin Company, 1988.

38. Huntsberger, David H. and Billingsley, Patrick. *Elements of Statistical Inference, 6th Ed*. Boston, Mass.: Allyn and Bacon, Inc., 1987.

39. Kelly, Louis G. *Handbook of Numerical Methods and Applications.* Reading, Mass.: Addison-Wesley Publishing Company, 1967.

40. Kennedy, Jr., William J. and Gentle, James E. *Statistical Computing*. New York: Marcel Dekker, Inc., 1980.

41. Kerlinger, Fred N. and Pedhazur, Elazar J. *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart and Winston, Inc., 1973.

42. Lieberman, Bernhardt (Editor). *Contemporary Problems in Statistics. A book of Readings for the Behavioral Sciences*. New York: Oxford University Press, 1971.

43. Lindgren, B. W. and McElrath, G. W. *Introduction to Probability and Statistics, 2nd Ed*. New York: The Macmillan Company, 1966.

44. Marcoulides, George A. and Schumacker, Randall E. (Editors). *Advanced Structural Equation Modeling. Issues and Techniques*. Mahwah, N.J.: Lawrence Erlbaum Associates, 1996.

45. Masters, Timothy. *Practical Neural Network Recipes in C++*. San Diego, Calif.: Morgan Kaufmann, 1993.

46. McNeil, Keith, Newman, Isadore and Kelly, Francis J. *Testing Research Hypotheses with the General Linear Model*. Carbondale, Ill.: Southern Illinois University Press, 1996.

47. McNemar, Quinn. *Psychological Statistics, 4th Ed*. New York: John Wiley and Sons, Inc., 1969.

48. Minium, Edward W. *Statistical Reasoning in Psychology and Education, 2nd Ed*. New York: John Wiley and Sons, Inc., 1978.

49. Montgomery, Douglas C. *Statistical Quality Control*. New York: John Wiley and Sons, Inc., 1985.

50. Mulaik, Stanley A. *The Foundations of Factor Analysis*. New York: McGraw-Hill Book Company, 1972.

51. Myers, Jerome L. *Fundamentals of Experimental Design*. Boston, Mass.: Allyn and Bacon, Inc., 1966.

52. Nunnally, Jum C. *Psychometric Theory*. New York: McGraw-Hill Book Company, Inc., 1967.

53. Olson, Chester L. *Essentials of Statistics. Making Sense of Data*. Boston, Mass.: Allyn and Bacon, Inc., 1987.

54. Payne, David A., Editor. *Curriculum Evaluation. Commentaries on Purpose, Process, Product*. Lexington, Mass.: D. C. Heath and Company, 1974.

55. Pedhazur, Elazar J. *Multiple Regression in Behavioral Research. Explanation and Prediction. 3rd Ed*. Fort Worth, Texas: Holt, Rinehart and Winston, Inc., 1997.

56. Press, William H., Flannery, Brian P., Teukolsky, Saul A., and Vetterling, William T. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, 1988.

57. Ralston, Anthony and Wilf, Herbert S. *Mathematical Methods for Digital Computers*. New York: John Wiley and Sons, Inc., 1966.

58. Rao, C. Radhakrishna. *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons, Inc., 1965.

59. Rao, Valluru and Rao, Hayagriva. *C++ Neural Networks and Fuzzy Logic, 2nd Ed*. New York: MIS Press, 1995.

60. Rogers, Joey. *Object-Oriented Neural Networks in C++*. San Diego, Calif.: Academic Press, 1997.

61. Roscoe, John T. *Fundamental Research Statistics for the Behavioral Sciences, 2nd Ed*. New York: Holt, Rinehart and Winston, Inc., 1975.

62. Rummel, R. J. *Applied Factor Analysis*. Evanston, Ill.: Northwestern University Press, 1970.

63. Scheffe', Henry. *The Analysis of Variance*. New York: John Wiley and Sons, Inc., 1959.

64. Schumacker, Randall E. and Lomax, Richard G. *A Beginner's Guide to Structural Equation Modeling*. Mahwah, N.J.: Lawrence Erlbaum Associates, 1996.

65. Siegel, Sidney. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill Book Company, Inc., 1956.

66. Silverman, Eliot N. and Brody, Linda A. *Statistics. A Common Sense Approach*. Boston, Mass.: Prindle, Weber and Schmidt, Inc., 1973.

67. SPSS, Inc. *SPSS-X User's Guide, 3rd Ed*. Chicago, Ill.: SPSS, Inc., 1988.

68. Steele, Sara M. *Contemporary Approaches to Program Evaluation: Implications for Evaluating Programs for Disadvantaged Adults*. Syracuse, New York: ERIC Clearinghouse on Adult Education (undated).

69. Stevens, James. *Applied Multivariate Statistics for the Social Sciences, 3rd Ed*. Mahwah, N.J.: Lawrence Erlbaum Associates, 1996.

70. Stodala, Quentin and Stordahl, Kalmer. *Basic Educational Tests and Measurement*. Chicago, Ill.: Science Research Associates, Inc., 1967.

71. Thomson, Godfrey. *The Factorial Analysis of Human Ability, 5th Ed.*, Boston, Mass.: Houghton Mifflin Company, 1951.

72. Thorndike, Robert L. *Applied Psychometrics*. Boston, Mass.: Houghton Mifflin Company, 1982.

73. Thorndike, Robert L. (Editor.) *Educational Measurement, 2nd Edition*. One Dupont Circle, Washington D.C.: American Council on Education, 1971.

74. Veldman, Donald J. *Fortran Programming for the Behavioral Sciences.* Holt, Rinehart and Winston, New York, 1967, pages 308-317.

75. Walker, Helen M. and Lev, Joseph. *Statistical Inference*. New York: Henry Holt and Company, 1953.

76. Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill Book Company, Inc., 1962.

77. Worthen, Blaine R. and Sanders, James R. *Educational Evaluation: Theory and Practice*. Belmont, Calif.: Wadsworth Publishing Company, Inc., 1973.

78. Yamane, Taro. *Mathematics for Economists*. *An Elementary Survey*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962.

Index