

System Overview

General

The Statistical Solutions System for Missing Data Analysis incorporates the latest advanced techniques for imputing missing values. Researchers are regularly confronted with the problem of missing data in data sets. According to the type of missing data in your data set, the Statistical Solutions System allows you to choose the most appropriate technique to apply to a subset of your data.

In addition to resolving the problem of missing data, the system incorporates a wide range of statistical tests and techniques. This manual discusses the main statistical tests and techniques available as output from the system. It is not a complete reference guide to the system.

About this Manual

Chapter 1

Data Management

Describes how to import a data file, specifying the attributes of a variable, transformations that can be performed on a variable, and defining design and longitudinal variables. Information about the Link Manager and its comprehensive set of tools for screening data is also given.

Chapter 2

Descriptive Statistics

Discusses Descriptive Statistics such as: Univariate Summary Statistics, Sample Size, Frequency and Proportion.

Chapter 3

Regression

Introduces Regression, discusses Simple Linear, and Multiple Regression techniques, and describes the available Output Options.

Chapter 4

Tables

(Frequency Analysis)

Introduces Frequency Analysis and describes the Tables, Measures, and Tests output options, which are available to the user when performing an analysis.

Chapter 5

t- Tests and

Nonparametric Tests

Describes Two-group, Paired, and One-Group *t*-Tests, and the Output Options available for each test type.

Chapter 6

Analysis (ANOVA)

Introduces the Analysis of Variance (ANOVA) method, describes the One-way ANOVA and Two-way ANOVA techniques, and describes the available Output Options.

Chapter 7

Plots

An overview of the plots available in the system, such as: Scatterplot, Histogram, Bar Chart and Normal Probability Plot.

Chapter 8

Tutorial

A tutorial with examples you can work through.

Appendices

Includes references, error messages and data set information.

Index

Index contents, subject matter, topics.

1. Data Management

SYSTEM PREFERENCES AND READING/SAVING DATA

SPECIFYING VARIABLE ATTRIBUTES

GROUPING VARIABLES

TRANSFORMING VARIABLES

DEFINING VARIABLES

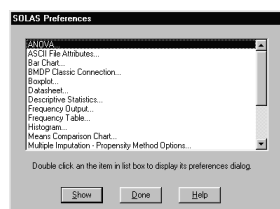
THE LINK MANAGER

Introduction

This chapter introduces you to setting preferences for your output options, importing data from other applications, the attributes of a variable, cutting and pasting variables and cases in the datasheet, transformations that can be performed on a variable, and defining design and longitudinal variables. Also there is an explanation of the Link Manager that comprises a powerful set of tools that you can use for screening data.

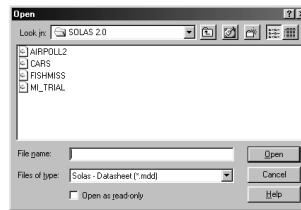
System Preferences

You can set preferences for the output options for an analysis by selecting the **View** menu **System Preferences** option in the Main window to display the window shown below:



Reading/Saving Data

To import a file into the system select **Open** from the pull-down **File** Menu. The Open window is displayed where the file to be opened can be selected.

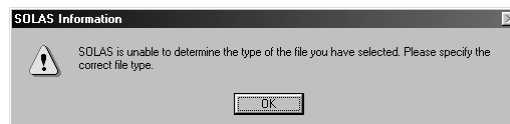


Supported File Types

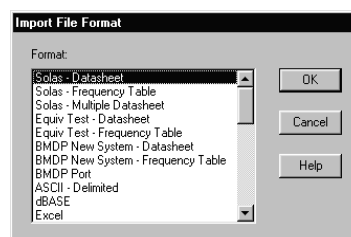
See the Readme.txt for the most up to-date list of supported file formats and also Appendix D: External File Formats. The system can read any of the following file formats:

Datasheet	Lotus 1-2-3 Worksheet
Frequency Table	Paradox File
Multiple Datasheet	Quattro Pro Worksheet
BMDP New System - Datasheet	SAS for Windows/OS2
BMDP New System - Frequency Table	SAS for Unix, SAS JMP
BMDP Port File	SAS Transport File
ASCII - Delimited	S-PLUS File
dBASE or compatible	SPSS Portable File
Excel Worksheet and Excel for Office 2000	SPSS Data File
FoxPro	Stata File
Minitab versions 8-12	Statistica version 5
Gauss File	SYSTAT File

If you enter a non-system file type in the **File Name** datafield, you will be prompted to specify a system file type:



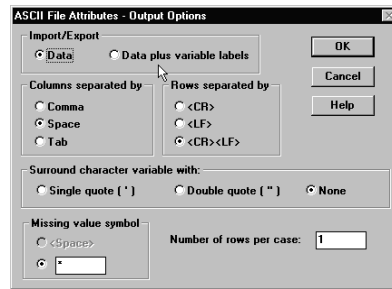
After pressing the **OK** button, the Import File Format window is displayed:



Select a file format and click **OK**. The system copies the selected file format into the datasheet.

Import an ASCII File

If the selected file type is ASCII the following window is displayed:



Import/Export

The **Import/Export** selection buttons allow you to specify whether the file you want to import contains Variable Names in the first record:

- Data (default)** The system reads the first record as variable data or values.
- Data plus Labels** The first record is read as Variable Names and all subsequent records as data.

Columns Separated By

The **Columns separated by** selection buttons allow you to choose the character that separates the variables within the record. You can choose among comma, space, or tab. The system defaults to space.

Rows Separated By

The **Rows separated by** selections buttons allow you to choose the end of record marker that separates records within a file. You can choose among carriage return <CR>, line feed <LF>, or both <CR><LF>. The system defaults to <CR><LF>.

Surround character variable with

The **Surround character variable with** selection buttons allow you to specify characters to be used as “surround” characters for your alphanumeric (character) variables. ASCII text files can have single quote, double quote, or no surround characters. The system defaults to none.

Missing Value Symbol

The **Missing Value Symbol** selection buttons allow you to specify the characters to be used in the datasheet to represent missing values. Only one missing value code is allowed when importing ASCII.

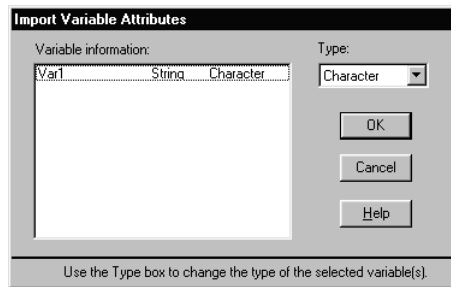
You can use a space as a character, but only when commas or tabs separate the data. You can also type in up to seven characters using any printing characters. The system defaults to an asterisk.

Number of Rows Per Case

The **Number of rows per case** datafield allows you to indicate the number of rows per case. For example, if each case takes up three rows of data, enter a 3 in the box. The system defaults to 1 row per case. Each case should have the fixed number of rows as specified.

Import Variable Attributes Window

The Import Variable Attributes window is displayed when you want to import any file including ASCII files. It displays a list of the variables in the file to be imported with their associated Statistical Solutions System type. You can change their type here before the file is imported into the system.



To change an imported variable type:

1. Import a file.
2. Highlight a variable(s) in the **Variable information** list, and select a new type from the **Type** datafield.
3. Click on **OK** when you are satisfied with all your choices.

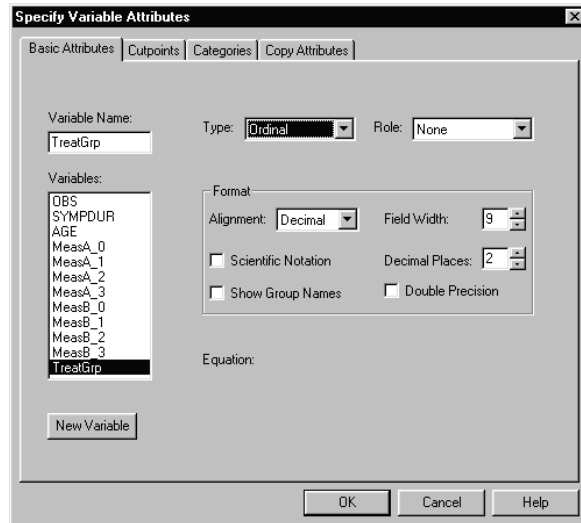
The system imports your file and displays it in a new datasheet.

NOTE: Generally the system does not handle date variables. When a date variable is detected, the system forces this variable to represent the number of days since Jan 1, 1900. You may also choose to read date variables as character strings or as continuous variables

Specifying Variable Attributes

In general, you will want to name your own variables rather than using default names Var1, Var2, etc. Most likely, you will also prefer to specify the type of data in the variables. You can name variables and specify variable type, role and format through the Specify Variable Attributes window.

To display the Specify Variable Attributes window, double click on the name of a variable in the datasheet. Alternatively, you can choose **Variable Attributes** from the **Variables** menu to display the Specify Variable Attributes window.



The Specify Variable Window is a tabbed window that lets you specify the attributes specific to each variable in your datasheet. It also lets you set cutpoints, modify categories and copy attributes. The four tabs are:

Basic Attributes	Specifies basic variable attributes.
Set Cutpoints	Groups continuous or ordinal variables.
Modify Categories	Groups nominal or ordinal variables.
Copy Attributes	Copies attributes from one variable to another variable(s).

Specify Basic Attributes

You can use the Basic Attributes window to change from the default Variable Name to any name not already assigned to a variable in the same datasheet. Each variable must be assigned a type. The type you assign to a variable affects its availability in certain plots and analyses.

A variable can be considered nominal, ordinal, continuous, integer or character. The default is continuous. The terminology used is comparable to the Stevens classification system (see Variable Type later in this section). The tabs on the window change according to the type of variable chosen.

Role

The Role of a variable can be: None, Grouping, Case Frequency, or Case Label. The default is None, which you should interpret as any.

Format

The Format section of the Basic Attributes window provides fields for specifying the format of the values in the variable when displayed in the datasheet.

Alignment

The Alignment field lets you choose among Left, Right, Center, and Decimal.

Field Width

The Field Width of a variable refers to the number of characters that fit into one cell of the datasheet. You may want to specify a field width larger than that required by the data in the variable when you have a long variable name.

Scientific Notation

You can specify whether or not your variables are to be displayed in Scientific Notation.

Decimal Places

You can specify the number of Decimal Places that are to be the values for each variable. Decimal Places has a default of 2, with a range from 0 to 8.

Show Group Names

Use the Show Group Names option to show group names rather than original values for nominal, ordinal, continuous and integer variables.

Double Precision

Double Precision is applicable to Continuous variables only. Double Precision uses a greater range and provides more accuracy.

New Variables

The New Variables option lets you create a new variable that is appended to the datasheet.

Append

You can also use the Append option in the Variables menu. The default type of a new variable is continuous.

Variable Type

The **Type** scrolled datafield displays four choices for your variable type. These are:

- ◆ Continuous
- ◆ Ordinal
- ◆ Nominal
- ◆ Integer.

These four choices are based on both the Stevens' classification system (ordinal and nominal) and the usual mathematical/ statistical definitions (continuous and integer.) The default value is continuous.

Continuous variable

Measurements that can take on any value (up to the limit of the accuracy of the measurement) within the possible range of that variable. Examples would be weight, height, or density. For such a variable, equal size differences along different parts of the scale are equivalent.

You can use continuous data as if it were nominal or ordinal by categorizing it using the cutpoint option to separate it into separate categories.

Integer variables

Any number such as 1, 2, 3, etc. or -1, -2, -3, etc. They will be stored without decimal points. They are often used for counted data. The program will allow you to do all available analyses with integer data. Statistics such as means computed from them can have decimal values.

Nominal variable

These are numerical representations of categories, or status. Examples are race, religion, or blood type. The categories are given numerical values but since the categories can be given any order, any set of numbers can be used to represent them. For example, the user might note the four blood types as 1= A, 2=B, 3=AB, and 4 = O but any other four distinct numbers could also be used. The concept of order has no meaning since 2 is not greater than 1, etc. Note that nominal variables can be sorted using the **Edit** menu **Sort** option.

If you classify a variable as nominal, that variable will not appear in lists of variables for statistics such as means or variances. You can use nominal variables as explanatory or independent variables in regression analysis but the system prompts you to create design variables for them. See Entering Variables Multiple Regression in Chapter 3.

Ordinal data

Used to imply an underlying order such as a wellness scale. Where a value of 1= poor, 2 = fair, 3 = good, and 4 = excellent. There is an underlying order or ranking to this data, but the difference between a rating of poor health and fair health may not be equal to the difference between good and excellent health even though they are both 1 unit apart. Ordinal values do not have to be integers.

A character value or name rather than a number may denote both nominal and ordinal variables. For example, male or female instead of 1 and 2 or mild, moderate and severe instead of 1,2, and 3 for an illness scale.

Data may be entered as character data initially, or the **Categories** tab can be used to associate a character value with each numeric value. To change from a number to a character value:

1. Double-click on a nominal or ordinal variable in the Datasheet.
2. Choose the **Categories** tab from the displayed Specify Variable Attributes window.
3. Type the desired name in the Group_Name field and press **Enter** after each entry.

Stevens also had interval and ratio variables. Here the continuous variable option should be used for those types of variables. If you classify the variables as ordinal, the program will allow you to do any available statistical analysis. Note that this is contrary to the advice of Stevens. Obviously you can call your data any of the four types of variables you wish. Classifying a variable as nominal is the only classification that restricts the use of the variable. See Velleman and Wilkinson (1993) for a discussion of why it is not useful to adhere too strictly to a classification scheme such as Stevens.

Variable Role

The Role scrolled datafield allows five options that affect how the variables are used in the planned analysis. The default role is None.

None

Signifies any role appropriate for the variable type. Other roles are Grouping, X Variable, Y Variable, Case Label, and Case Frequency.

Case Label

Variables are used to identify individual cases. Hence, for a variable to be useful as a case label, it must have a unique value for each case. In a given datasheet, only one variable can have the role Case Label at any given time. If you specify a case label variable, it will be used as a label wherever applicable (e.g., scatterplot, missing data pattern) without your being asked. Case labels may assume other roles in an analysis, but actual case labels will rarely be appropriate for anything else (because they have a unique value for each case).

Aside from case label, the role of a variable does not determine its use. A variable with role grouping, x variable, y variable and case frequency may be used in any role for which it is otherwise appropriate.

Case Frequency

Variables act as weights. The weight has the effect of repeating the case, although cases are not actually added to the list. If case one has a weight of 3 and case two has a weight of 5, then the system computes the mean of these two cases as $(3X_1 + 5X_2)/8$.

Case frequencies are used when like cases are lumped together, or when some respondents do not respond in any way to a survey, but there is some information available on them from the sampling design. In a given datasheet, at most one variable can have the role Case Frequency at any given time.

X and Y Variables

If your datasheet contains a case frequency variable, each time you select an analysis, you will be asked whether the variable should be used as a frequency variable. If you click **Yes**, that variable will not be available for use as an x, y, or grouping variable in the analysis. If you click **No**, it will be available for any role for which it is otherwise appropriate.

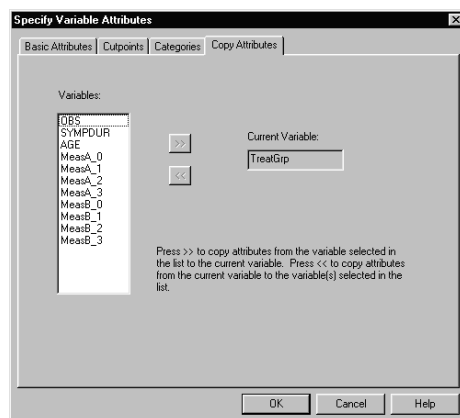
If you have a datasheet with designated roles that exactly match requirements for a given analysis (e.g., a y variable and two grouping variables for a 2-way ANOVA), the specification dialog will open with the likely choices.

If there is not an exact match with the requirements of the analysis (e.g., you have two y variables and 2 grouping variables), the dialog will not make any guesses for you. Receiving suggested values in specification dialogs is the only affect of setting x and y variables in your datasheet. See Grouping Variables for a discussion of this role.

Copy Attributes

Select the Copy Attributes tab from the Specify Variable Attributes window. The variable in the Variable Attributes window will always be either the *source* or the *recipient* of the attributes.

1. Press the right-facing arrows button if you want to copy attributes from the variable selected in the Variables list to the variable in the Current Variable datafield.
2. Press the left-facing arrows button if you want to copy attributes from the variable in the Current Variable datafield to the variable(s) selected in the Variables list.



If you created a new variable that currently contains no data values, you are still able to copy attributes to it.

Cutting Datasheet Variables and Cases

To cut variables or cases from a datasheet:

1. Select a variable from the datasheet by clicking on its name. A vertical column becomes highlighted.
2. Select a case by clicking on the case ID area. A horizontal row becomes highlighted.

Click on **Cut** in the **Edit** Menu. The variable or case becomes removed and placed on the clipboard.

	Name	Rain	Education	Pop_den	Nonwhite	Nox
1	akronOH	36	11.4	3243	8.8	15
2	albanyNY	35	11.0	4281	3.5	10
3	allenPA	44	9.8	4260	0.8	6
4	atlantGA	47	11.1	3125	27.1	8
5	baltimMD	43	9.6	6441	24.4	38
6	birmnghAL	53	10.2	3325	38.5	32
7	bostonMA	43	12.1	4679	3.5	32
8	bridgeCT	45	10.6	2140	5.3	4
9	buffaloNY	36	10.5	6582	8.1	12
10	cantonOH	36	10.7	4213	6.7	7
11	chatagTN	52	9.6	2302	22.2	8
12	chicagIL	33	10.9	6122	16.3	63
13	cinncoOH	40	10.2	4101	13.0	26
14	clevelandOH	35	11.1	3042	14.7	21

Pasting Datasheet Variables and Cases

To paste a variable or case back into the datasheet by placing an insertion line to indicate the position for the variable or case:

1. Place an insertion line between two variables by clicking on the line in the variable name area between the two variables.
2. Place an insertion line between two cases by clicking on the line in the case name area between the two cases. The line becomes thicker than the other vertical or horizontal lines.
3. Click on **Paste** in the **Edit** menu to paste the variable or case.

	Name	Rain	Pop_den	Nonwhite	Nox	So2	Mortality
1	akronOH	36	3243	8.8	15	59	921.9
2	albanyNY	35	4281	3.5	10	39	997.9
3	allenPA	44	4260	0.8	6	33	962.4
4	atlantGA	47	3125	27.1	8	24	982.3
5	baltimMD	43	6441	24.4	38	206	1071.0
6	birmnghAL	53	3325	38.5	32	72	1030.0
7	bostonMA	43	4679	3.5	32	62	934.7
8	bridgeCT	45	2140	5.3	4	4	899.5
9	buffaloNY	36	6582	8.1	12	37	1002.0
10	cantonOH	36	4213	6.7	7	20	912.3
11	chatagTN	52	2302	22.2	8	27	1010.0
12	chicagIL	33	6122	16.3	63	278	1025.0
13	cinncoOH	40	4101	13.0	26	146	970.5
14	clevelandOH	35	3042	14.7	21	64	986.0
15	columbOH	37	4259	13.1	9	15	958.8
16	dallasTX	35	1441	14.8	1	1	860.1
17	daytonOH	36	4029	12.4	4	16	936.2
18	denverCO	15	4824	4.7	8	28	871.8
19	detrotMI	31	4834	15.8	35	124	959.2

Grouping variables

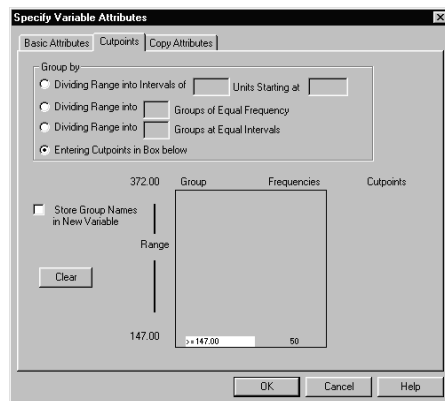
When you want to use a continuous or an ordinal variable as a grouping variable, you must define groups by setting cutpoints for that variable.

Set Cutpoints

If you choose to categorize continuous or ordinal data, you will use the graphical Cutpoints window. In this window you can see the range of your measurements for the chosen variable. You can easily add cutpoints one at a time, and you can drag any cutpoint to a different position.

To specify cutpoints for a grouping variable:

1. Start from the datasheet. Choose **Group** and **Set Cutpoints** from the **Variables** menu, or choose the Cutpoints tab from the Specify Variable Attributes window. If there are no ordinal or continuous variables in the datasheet, the Cutpoints option will be grayed out.



2. In the Cutpoints window, click on the variable to be grouped. The window provides four options. Most of the time you will find that choosing one of the first three options will work well for you. Click on your choice, then fill in the blanks. Place the cursor in the appropriate field, and type in a numerical value.
3. If you prefer to work graphically, you can enter cutpoints in the large cutpoint box. Click on the button to choose the Entering Cutpoints in Box below option. Place the cursor in the large cutpoint box and click. A line will appear.
4. The first cutpoint line separates the sample into two groups. For each group, the box displays the range and the frequency of cases in the group. You can drag the line up or down and click each time that you want to put in an additional line.
5. The graphical cutpoint box in the system gives you complete freedom to choose the exact cutpoints that you want. The program also lets you see the results instantaneously. If you do not like your choice, click on Clear and try again.

- When you are finished, click **OK**.

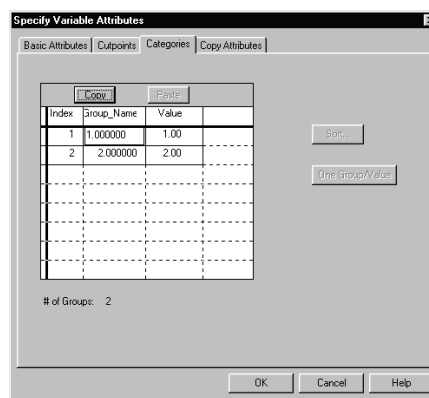
To view the descriptive statistics output with your new grouping:

- Choose the **Descriptive Statistics** option from the **Analyze** menu, and then select **Continuous/Ordinal**. The Descriptive Statistics window will appear without grouping.
- To add grouping, choose the **Grouping** option from the **Options** menu, and select your new grouping variable.

Modify Categories

Select the Categories tab from the Specify Variable Attributes window or choose **Group** and **Modify Categories** in the Datasheet **Variables** Menu. The system displays the Categories window.

Only the following tabs are displayed when there are no nominal or ordinal variables in a datasheet: Basic Attribute, Cutpoints and Copy Attributes. New category names can be entered in the Group Name field.



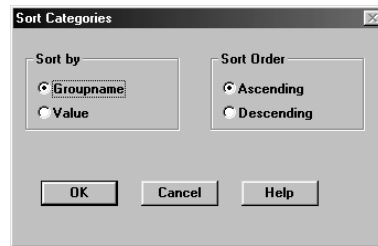
Reorder Categories

Clicking to the left of the bold line of the Index column selects the desired category. The category becomes highlighted and the **Copy** button at the top of the screen changes to **Cut**. When you click on the **Cut** button, the category is removed.

Position the cursor at the junction of the bold line and a horizontal line marking the position for pasting the cut category, the line darkens. Clicking on the **Paste** button causes the category to be inserted in the specified position.

You cannot reorder ordinal variables stored as integers. If you reorder groups for more than one variable without leaving the window between the changes, the system asks you if you want to save the displayed categories.

You can use the Group Name field to sort categories by name, and the Value field to sort categories by value. Specify Ascending/Descending for the order in which you want to sort. The **One Group/Value** button lets you specify one group for each value. Pressing the **Sort** button displays the Sort Categories window.



Transforming Variables

The system allows you to create new variables by transforming existing variables in your datasheet. When you define your own transformation you are writing a StatSol Scripting Language (SSL) expression. If you write an expression that is not understood, the system displays an error message detailing the input expected by the system. The list of errors is given in Appendix C, Error Messages.

You can transform a selected variable in your datasheet from the menu or you can define a transformation with the **User-defined transformation** option. This option allows you to define new variables based on more than one variable and/or more than one function. It also provides an extended set of functions from which to select. When you transform a variable in your datasheet. The transformation affects all cases of the variable.

Operators

Use arithmetic and conditional Operators to write a StatSol Scripting Language (SSL) expression. Be sure to include a space on either side of the Not, And, and Or operators. You may also type in the formula directly, or edit an existing formula. You can enter keyboard symbols and numbers directly for addition, multiplication, etc. A table showing the operators, their meaning and their keyboard entries is given below:

Operator	Meaning	Keyboard Entry
+ or -	Addition and subtraction	+ or -
*	multiplication	*
/	division	/
$a**b$	exponentiation	**
=	equal to	=
≠	not equal to	/=
>	greater than	>
≥	greater than or equal to	>=
<	less than	<
≤	less than or equal to	<=
Not	not	Not
And	and	And
Or	or	Or

Pre-Defined Functions

The table below shows the pre-defined functions available for user-defined transformations:

Groupings	Functions
Common Transformations	log, ln, sqrt, exp(e**X), abs, sign, sq(X**2), inv(1/x), int, mod(x modulus(p)), (X+a)**p, BoxCox
Trigonometric Functions	sin, cos, tan, arcsin, arccos, arctan
Summary Functions	N, NMissing, Mean, Median, StdDev, sum, min/max, intercept, slope, r, area, trapArea, trend, trendConst, correlation
Cumulative Distribution Functions	Normal, Student- <i>t</i> , Chi-square, <i>F</i> , Inverse Normal, Inverse Student- <i>t</i> , Inverse Chi-square, Inverse <i>F</i>
Multipass Functions	Lag, deviate, absdeviation, z score, trimmed, winsorized
Frequency Distribution Functions	Normal, Uniform

Transform Window

The Transform window has four tabs which when pressed display the following views:

- ◆ Common
- ◆ Trigonometric
- ◆ Summary
- ◆ Multipass

If the Transform window is NOT displayed, then the following two notes apply:

NOTE 1: If you select (highlight) a variable column in your datasheet, and then apply a simple transform from the displayed **Variables → Transform** menu, the system inserts the new variable to the right of the selected variable in your datasheet.

NOTE 2: If you select (highlight) a vertical line in any column in your datasheet, and then apply a simple transform from the displayed **Variables → Transform** menu, the new variable is inserted at that point.

Otherwise, the transformed variable is displayed in the column to the right of the last variable entered in your datasheet.

NOTE: Transformations require matching parentheses in the expression.

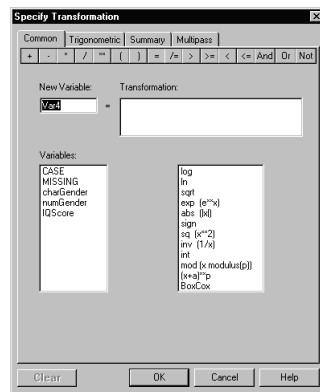
Common and Trigonometric Transformations

You can use Trigonometric functions (sin, cos, tan, arcsin, arccos, arctan) when a variable is an angle and you want the sine, cosine, or tangent of the angle. You must provide the angle in radians.

Degrees to Radians To change an angle given in degrees to radians, multiply by $\pi/180=0.017453$.

Inverse trigonometric functions are arc sines, arc cosines, and arc tangents. The system gives the result in radians. The arc sine transformation is sometimes used with binomial data.

From the Variables menu, select **Transform** → **User-defined transformation**, the Transform window is displayed.



Common transformation

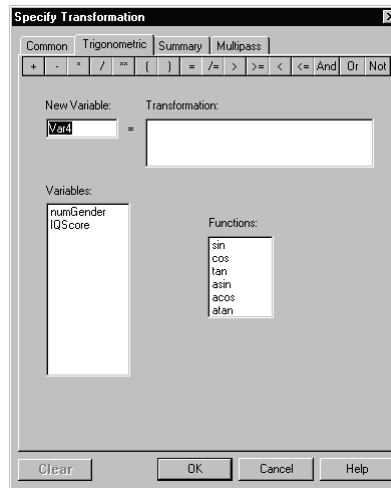
To perform a Common transformation, execute the following steps:

1. Click on the variable to be transformed in the Variables listbox.
2. The selected variable is displayed in the Transformation datafield.
3. Select a common function from the listbox, or an arithmetic operator from the selection buttons at the top of the window. The expression is displayed in the Transformation datafield.
4. When the transformation expression is complete, you can modify the variable name in the New Variables datafield, or accept the default name.
5. Press the **OK** button to perform the transformation and display the transformed variable in the datasheet, or press the **Clear** button to redo the transformation.

Trigonometric transformation

To perform a trigonometric transformation, execute the following steps:

1. Press the Trigonometric tab in the Specify Transformation window to display the Trigonometric view as shown below:
2. Click on the variable to be transformed in the Variables listbox.
3. The selected variable is displayed in the Transformation datafield.
4. Select a trigonometric function from the listbox, or an arithmetic operator from the selection buttons at the top of the window. The expression is displayed in the Transformation datafield.
5. When the transformation expression is complete, you can modify the variable name in the New Variables datafield, or accept the default name.
6. Press the **OK** button to perform the transformation and display the transformed variable in the datasheet, or press the **Clear** button to redo the transformation.



Common Transformations Definitions

log

The log function returns the logarithm of a variable (base 10). The system language representation is:

transformed variable = log(original variable)

ln

The ln function returns the logarithm of a variable to the base e . The system language representation is:

transformed variable = ln(original variable)

sqrt

The sqrt function returns the square root of a variable. The system language representation is:

transformed variable = sqrt(original variable)

exp(e**X)

The exp function is an exponential transformation on the original variable X and returns the mathematical constant e (≈ 2.71828) raised to a power given by X . The system language representation is:

transformed variable = exp(original variable)

This results in large numbers if X is large; do not use if the original value is greater than 80.

abs

The abs function returns the absolute value of a variable. The transformed variable is equal to the original variable X if X is non-negative, or equal to the negative of X if X is negative. The system language representation is:

transformed variable = abs(original variable)

sign

The sign function of the original variable X returns a value of: -1 if X is negative, and 1 if X is zero or positive. The system language representation is:

transformed variable = sign(original variable)

sq (X^{2})**

The sq function returns the square of the original variable X , or X times X . The system language representation is:

transformed variable = sq(original variable)

inv (1/ X)

The inv function returns the inverse of the original variable X , or $1/X$. The system language representation is:

transformed variable = inv(original variable)

int

The int function returns the integer part of a variable. The transformed variable is equal to the original variable X if X is an integer; equal to the largest integer that is less than X if X is positive; or equal to the smallest integer that is greater than X if X is negative. The system language representation is:

transformed variable = int(original variable)

mod ($x \bmod p$)

The mod function returns the remainder of the original variable X divided by p . The system language representation is:

transformed variable = mod(original variable, p)

where p is a positive number.

When you choose this function, the system displays a template. Drag and drop the appropriate variable into the x field. Then enter a value for p .

($X+a$) **p

The function $(X+a)^{**p}$ is a widely used power function. Commonly-chosen values of p are -1, -.5, .5, or 2. The system language representation is:

transformed variable = (original variable + a) **p

Where the constant a is commonly taken as zero, but can not be less than, or equal to, the negative of the minimum value of X for some values of p . For example, you cannot take the square root of a negative number, or missing values will be generated. The $(X+a)^p$ option lets you define one or more power transformations of a given variable.

When you choose this function, the system displays a template.

Drag and drop the appropriate variable into the x field. Then enter a value for a and a value for p .

BoxCox

The BoxCox function returns the following transformations:

$$\text{transformed variable} = ((\text{original variable} + a)^p - 1)/p \quad p \neq 0$$

$$\text{transformed variable} = \ln(\text{original variable} + a) \quad p = 0$$

where a is a constant and p is any number. The system language representation is:

$$\text{transformed variable} = \text{BoxCox}(\text{original variable}, a, p)$$

When you choose this function, the system displays a template. Drag and drop the appropriate variable into the X field. Then enter a value for a and a value for p .

Trigonometric Transformations Definitions**sin**

The sin function returns the trigonometric sine of the original variable X , where X is assumed to be in radians. The system language representation is:

$$\text{transformed variable} = \sin(\text{original variable})$$

cos

The cos function returns the trigonometric cosine of the original variable X , where X is assumed to be in radians. The system language representation is:

$$\text{transformed variable} = \cos(\text{original variable})$$

tan

The tan function returns the trigonometric tangent of the original variable X , where X is assumed to be in radians. The system language representation is:

$$\text{transformed variable} = \tan(\text{original variable})$$

arcsin

The arcsin function is the inverse trigonometric function of the sine and returns a value in radians. The absolute value of the original variable X must be less than or equal to 1. The system language representation is:

$$\text{transformed variable} = \arcsin(\text{original variable})$$

arccos

The arccos function is the inverse trigonometric function of the cosine and returns a value in radians. The absolute value of the original variable X must be less than or equal to 1. The system language representation is:

$$\text{transformed variable} = \arccos(\text{original variable})$$

arctan

The arctan function is the inverse trigonometric function of the tangent in radians. The system language representation is:

$$\text{transformed variable} = \arctan(\text{original variable})$$

Summary Transformations

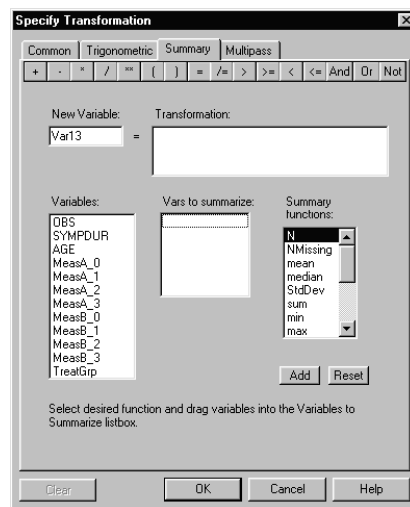
The **Summary** option in the Specify Transformation window allows you to create a new variable that summarizes data from two or more related variables (e.g., averages). The **Summary** option computes summary statistics based on a set of variables for each case. These include the number of valid values in the set (*N*), the number of missing values in the set (*NMissing*), the mean, median, standard deviation, sum, minimum value, and maximum value of the set.

The Summary view in the Specify Transformations window displays a listbox with options: *N*, *NMissing*, Mean, Median, StdDev, sum, min/max.

Summary transformation

To perform a Summary transformation, execute the following steps:

1. Press the Summary tab in the Specify Transformation window to display the Summary view as shown below:



2. Drag and drop the variable(s) to be summarized from the Variables listbox to the Variables to summarize listbox.

NOTE: If you wish to re-select different variables, press the **Reset** button to clear the Variables to summarize listbox.

3. Highlight a function in the Summary functions listbox, and press the **Add** button.

The selected variable and the function to be applied is displayed in the Transformation datafield.

When the transformation expression is complete, you can modify the variable name in the Name Variables datafield, or accept the default name.

4. Press the **OK** button to perform the transformation and display the transformed variable in the datasheet, or press the **Clear** button to redo the transformation.

Summary Functions Definitions

N

The *N* function returns the number of valid values in a set of variables for each case. The system language representation is: transformed variable = *N*(variable list), where variable list enumerates the names of variables in the set, separated by commas.

NMissing

The NMissing function returns the number of missing values in a set of variables for each case. The system language representation is: transformed variable = NMissing(variable list) where variable list enumerates the names of variables in the set, separated by commas.

mean

The mean function returns the average value in a set of variables for each case. The system language representation is:

transformed variable = mean(variable list)

where variable list enumerates the names of variables in the set, separated by commas.

See mean under Descriptive Statistics.

median

The median function returns the median value in a set of variables for each case. The system language representation is:

transformed variable = median(variable list)

where variable list enumerates the names of variables in the set, separated by commas.

StdDev

The StdDev function returns the standard deviation in a set of variables for each case. The system language representation is:

transformed variable = StdDev(variable list)

where variable list enumerates the names of variables in the set, separated by commas. At least two variables are required.

sum

The sum function returns the sum of values in a set of variables for each case. The system language representation is:

transformed variable = sum(variable list)

where variable list enumerates the names of variables in the set, separated by commas.

min/max

The min/max function returns the minimum/maximum value in a set of variables for each case. The system language representation is: transformed variable = min(variable list) , or, transformed variable = max(variable list), where variable list enumerates the names of variables in the set, separated by commas.

intercept

The intercept function returns the constant a of the regression line $y = a + bx$

slope

The slope function returns the coefficient b of the regression line $y = a + bx$

r

The r function returns the correlation coefficient $r(x,y)$

area

The area function returns the area under y : $(y_1 + 2 \cdot y_2 + \dots + 2 \cdot y_{n-1} + y_n) / 2$

trapArea

The trapArea function returns the area under y via the trapezoidal rule: $\text{sum}((x - x_{i-1}) \cdot (y_{i-1} + y_i) / 2)$

trend

The trend function returns the trend of (y_1, y_2, \dots, y_n) over $(1, 2, \dots, n)$

trendConst

The trendConst function returns the intercept of the trend line; i.e., constant of regression line through:

$(1, y_1), \dots, (n, y_n)$

correlation

The correlation function returns the correlation of $(1, 2, \dots, n)$ and (y_1, y_2, \dots, y_n)

Multipass Transformations

When you enter most transformations, you can simply choose one transformation after another to be added to your expression. However, there are a few transformations that cannot become part of an expression. If you try to use one of them in an expression, the system will give you a message stating that you cannot use it in an expression. For multipass transformations, the following cannot become part of an expression:

deviations, absolute deviation, z-scores, trimmed, winsorized, lag.

Multipass Functions

The Multipass option includes transformations of the variable X that require the system to go through the datasheet more than once. The variables formed by the multipass transformations differ from other transformations in their linkage: See the topic Linkage in the online help system. The resulting multipass-transformed variable is not linked to the original variable.

You **cannot** use a multipass transformation in a transformation with other variables or functions.

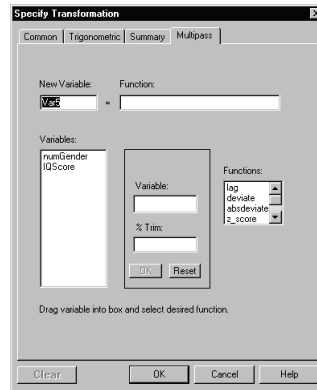
You **can** use a multipass-transformed variable in a subsequent transformation to create a new variable. If you do that, the two variables will be linked.

The definitions for the Multipass functions are given below.

Multipass transformation

To perform a Multipass transformation, execute the following steps:

1. Press the Multipass tab in the Specify Transformation window to display the Multipass view as shown below:



2. Drag and drop the variable for the Multipass transformation to the Variable datafield.
3. Select a function for the transform from the Functions scrolled listbox.

NOTE: If the Lag function is selected, the n datafield is displayed and a Lag value should be entered. If the Trimmed, or Winsorized functions are selected, a %Trimming datafield is displayed where a value should be entered.

4. Press the centre **OK** button to display the transform in the Function datafield, or press the **Reset** button to select a different variable for the transformation.
5. Press the lower **OK** button to perform the transformation and display the transformed variable in the datasheet, or press the **Clear** button to redo the transformation.

Multipass function definitions

deviations

The deviations function returns the deviation of the original variable X from its mean (i.e., X minus the mean of X). The system language representation is:

transformed variable = deviation(original variable)

absolute deviation

The absolute deviation function returns the absolute value of the deviation of the original variable X from its median (i.e., $|X$ minus the mean of $X|$). The system language representation is:

transformed variable = absdeviation(original variable)

z-scores

The `z-scores` function returns the standardized value of the original variable X (i.e., the deviation of X from its mean divided by its standard deviation). The system language representation is:

transformed variable = `z_score(original variable)`

lag

The `lag` function returns the lagged values of the original variable X . The default is of order one. That is, the i th value of the transformed variable is the $(i - 1)$ th value of the original variable. Under user-defined transformation, you can specify a general lag function of any order less than the number of cases, positive or negative. The system language representation is: transformed variable = `lag(original variable, n)`, where n is integer-valued. In general, the i th value of the transformed variable is equal to the $(i - n)$ th value of the original variable.

Missing values are generated when $(i - n)$ is less than 1 or greater than the total sample size N .

When you choose this function, the system displays a template. Drag and drop the appropriate variable into the `x` field. Then enter a value for n .

trimmed

The `trimmed` function returns:

- ◆ the values of the original variable X if the values fall in the region located in the middle $(1 - 2p)p$ percent of the data,
- ◆ missing values if X falls in the upper or lower $100p$ percent region.

The trimming region is determined by specifying a trimming percent level. The specified percentage of the sample is trimmed at both ends. Default is 5 percent. The trimmed values are the sample of X values that remains after discarding the i th largest and i th smallest values. The system language representation is:

transformed variable = `trimmed(original variable, p)`

where p is the trimming percent level; p should be between 0 and 0.25.

winsorized

The `winsorized` function returns:

- ◆ the values of the original variable X if the values fall in the region located in the middle $(1 - 2p)100$ percent of the data;
- ◆ the smallest X value in the middle region if the original value is in the lower $100p$ percent region; and
- ◆ the largest X value in the middle region if the original value is in the upper $100p$ percent region.

This function basically replaces extreme values with less extreme values in the data. The system language representation is:

transformed variable = winsorized(original variable, p)

where p is the trimming percent level; p should be between 0 and 0.25.

For an example using Multiple Transformation Functions, see the online help.

Cumulative and Frequency Distributions

Selecting the **Distributions** option in a datasheet **Variables** menu displays the Specify Distribution window that allows you to create a new variable and specify its Cumulative or Frequency distribution. This window has two selectable tabs, Cumulative Distributions and Frequency Distributions.

The Cumulative Distributions view has a listbox with the following selectable functions:

- ◆ Normal, Student- t , Chi-square, F , Inverse Normal
- ◆ Inverse Student- t , Inverse Chi-square, Inverse F .

The Frequency Distribution option allows you to generate one or more random samples that follow either the normal distribution or the uniform distribution.

The seed is an integer that is used by the random number generator. If you want to obtain a sample containing the same sequence and values in it as a previous sample, use as a seed the identical number that was used in the first sample. Otherwise, use a different seed. Suitable values for the seed range from 1 to 30,000.

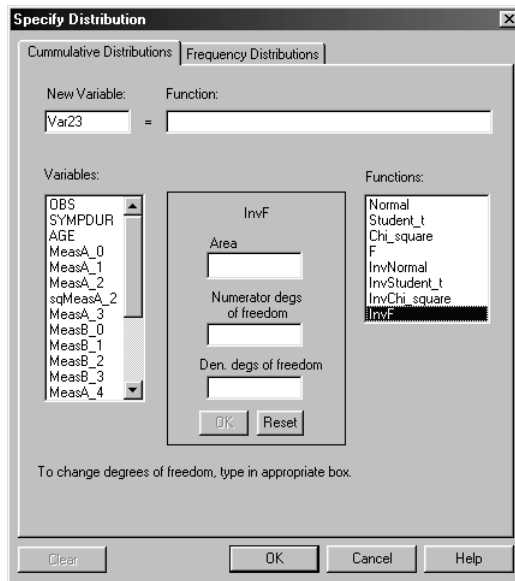
The Frequency Distributions view has a listbox with the following selectable functions:

- Normal or Uniform.

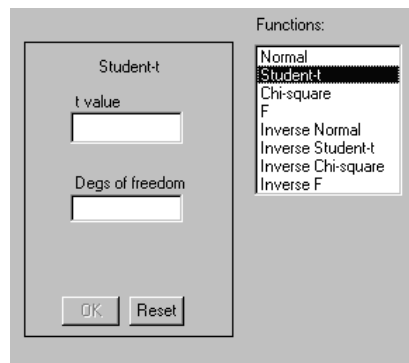
Cumulative Distribution

To perform a Cumulative Distribution transform on a variable, execute the following steps:

1. In a datasheet **Variables** menu select **Distributions** to display the Specify Distribution window as shown below:
2. Drag and drop a variable from the Variables listbox to the Value datafield.
3. Select the function that is to be applied to the variable.



NOTE: Depending on the function selected, additional datafields will be displayed as shown below:



The datafield shown above will be displayed when the following functions are selected: Student-t, Chi-square, Inverse student-t and Inverse Chi-square.

4. Enter an integer value for the Degrees of freedom.
5. If the values need to be modified at this stage, press the **Reset** button and enter the new values.
6. Press the center **OK** button. The expression is displayed in the Function datafield.
7. If the expression needs to be modified, press the **Clear** button and repeat from step 2.
8. When the expression in the Function datafield is correct, press the lower **OK** button, and the transformed variable will be displayed in the datasheet.

The datafield shown below will be displayed when the functions F and Inverse F are selected:

For the F and Inverse F functions, integer values should be entered in the Numerator and Denominator datafields, then continue from step 4 above.

Frequency Distribution

1. In a Specify Distribution window, press the Frequency Distribution tab to display the window shown below:

NOTE: Depending on the function selected, additional datafields will be displayed as shown below

normal

Mean:
0

Standard deviation:
1

Seed:
12315

OK Reset

2. Enter the Standard deviation and Seed values.
3. If the values need to be modified at this stage, press the **Reset** button and enter the new values.
4. Press the center **OK** button. The expression is displayed in the Function datafield.
5. If the expression need to be modified, press the **Clear** button and repeat from step 2.
6. When the expression is correct, press the lower **OK** button and the transformed variable will be displayed in the datasheet.

The datafield areas shown below will be displayed when the Uniform function is selected from the Specify Distribution window:

uniform

Minimum:
0

Maximum:
1

Seed:
12315

OK Reset

For the Uniform function, integer values should be entered in the Minimum and Seed datafields, then continue from step 3 above.

Cumulative Distribution Function Definitions

The following functions are defined as Cumulative Distributions:

Normal

The GCDF function returns the density area to the left of the original variable X based on a standard normal distribution.

The system language representation is:

transformed variable = GCDF(original variable).

Student-*t*

The TCDF function returns the density area to the left of the original variable X based on a Student's t distribution. The system language representation is:

transformed variable = TCDF(original variable, df)

where df is the degrees of freedom.

Chi-square

The CCDF function returns the density area to the left of the original variable X based on a chi-square distribution. The system language representation is:

transformed variable = CCDF(original variable, df)

where df is the degrees of freedom.

F

The FCDF function returns the density area to the left of the original variable X based on an F distribution. The system language representation is:

transformed variable = FCDF(original variable, $df1$, $df2$)

where $df1$ and $df2$ are the numerator and denominator degrees of freedom, respectively.

Inverse Normal

The GCDI function is the inverse function of GCDF and returns a quantile value corresponding to an area to the left given by the original variable X based on a standard normal distribution.

The system language representation is:

transformed variable = GCDI(original variable)

The value of X must be greater than zero and less than 1.

Inverse Student-*t*

The TCDI function is the inverse function of TCDF and returns a quantile value corresponding to an area to the left given by the original variable X based on a Student's t distribution. The system language representation is:

transformed variable = TCDI(original variable, df)

where df is the degrees of freedom. The value of X must be greater than zero and less than 1.

Inverse Chi-square

The CCDI function is the inverse function of CCDF and returns a quantile value corresponding to an area to the left given by the original variable X based on a chi-square distribution. The system language representation is:

transformed variable = CCDI(original variable, df)

where df is the degrees of freedom. The value of X must be greater than zero and less than 1.

Inverse F

The FCDI function is the inverse function of FCDF and returns a quantile value corresponding to an area to the left given by the original variable X based on an F distribution. The system language representation is:

transformed variable = FCDI(original variable, $df1$, $df2$)

where $df1$ and $df2$ are the numerator and denominator degrees of freedom, respectively. The value of X must be greater than zero and less than 1.

For an example using Cumulative Distributions, see the online help.

Frequency Distribution Function Definitions

Normal

The normal function returns a normal random deviate. The system language representation is:

transformed variable = normal(m , s)

where m and s are the population mean and standard deviation, respectively.

Uniform

The uniform function returns a uniform random deviate. The system language representation is:

transformed variable = uniform(min , max)

where min and max are the population minimum and maximum values.

For an example of using Random Number Functions, see the online help.

Defining Variables

Variables are entered and removed from analyses as a unit. Longitudinal Variables are derived from existing variable - related measures of some kind (e.g. blood pressure readings by week). In the case of missing data, these variables are used in the system to impute missing values using different techniques such as Multiple Imputation, Group Mean, Last Value Carried Forward and Hot Decking.

You can define variables by selecting the **Variables** menu **Define Variables** option and choosing **Design Variables** or **Longitudinal**, depending on the type of variable you wish to define. Windows are displayed where it is possible to define the Longitudinal variables and Design variables, and modify and remove existing variables.

Defining Longitudinal Variables

Longitudinal variables are variables which are intended to be measured at several points in time, for example, pre and post test, measurements of an outcome variable made at monthly intervals, laboratory tests made at each visit from baseline, through the treatment period, and through the follow-up period. If, for example, patients' blood pressures are to be recorded every month over a period of six months, we would refer to this as one longitudinal variable consisting of six repeated measures or periods.

The Define Longitudinal Variable window lists all currently defined Longitudinal Variables. The Type and Role of the current variable are indicated in the Type and Role fields. See definitions earlier in the chapter.

Longitudinal Variables

The Longitudinal Variables listbox contains any previously defined longitudinal variables. When you select a Longitudinal Variable name in the Variables listbox, the name of the Longitudinal Variable is placed in the Name field and the variable members appear in the Repeated Measurements list.

Name

The Name field contains a default name or the name of the Longitudinal Variable selected.

Variables

The Variables scrolled listbox shows all the variables in the data set, omitting only character variables and variables with no data.

Elements in Variable

The Elements in Variable table displays the current members of the current set of variables defining the longitudinal variable. You can drag variables into the Elements in Variable table from the Variables listbox. You can drag variables out of the Elements in Variable table to the Variables listbox.

New Variable button

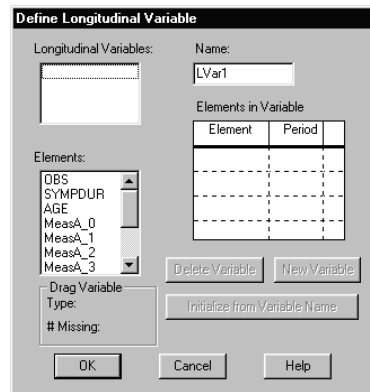
The New Variable button is enabled only when there are variables in Repeated Measurements list. It is used to define new variable sets.

Delete Variable

Use the Delete Variable button to remove a currently specified set. This button is enabled whenever there are elements in the Repeated Measurements list.

Initialize from Variable Name

The Initialize from Variable Name button is used to simplify the definition of set names. The Initialize from Variable Name button is enabled only if the initial or first part of a variable name matching a variable in the Variables listbox is entered into the Name field and you press return.



To define longitudinal variables, perform the following steps:

1. Open the file MI_TRIAL.MDD (located in the SAMPLES subdirectory), select the **Variables** menu **Define Variables** option and the **Longitudinal** option. This data set contains two longitudinal variables, each of which is made up of a baseline period and 3 post-baseline periods. The system assigns a default name of LVar1 to the first longitudinal variable.
2. Change this name to MeasA by typing the name into the Name field.
3. Click and drag the baseline variable (MeasA_0) from the Variables listbox to the first row of the Repeated Measurements field. The system automatically assigns a period value of zero to this element.
4. Drag the first post-baseline variable (MeasA_1), then the second post-baseline variable (MeasA_2), and finally the third post-baseline variable (MeasA_3). These elements will be assigned period values of 1, 2, etc. but you can change these by typing in new values.
5. For example, you might want to change the default period values if your repeated measurements were taken at month1, month6, and month8 i.e., at unequal time intervals. By setting the period values to 1, 6, and 8, you can ensure that linear interpolation of bounded missings will be correct. In this example, the measurements were taken at month1, month2, and month3, so we can keep the default values.
6. Click on New Variable to define the elements of our second longitudinal variable. A window is displayed asking if you want to save your changes to the longitudinal variable MeasA.
7. Click **Yes**.
8. Type the name MeasB in the name field.
9. Drag the baseline variable (MeasB_0) from the Variables listbox to the next blank row of the Repeated Measurements field.

10. Drag the first post-baseline variable (MeasB_1) from the Variables listbox to the next blank row of the Repeated Measurements field.
11. Continue dragging variables until you have defined all of the periods.
12. When you are satisfied that you have defined your longitudinal variable(s) correctly, click **OK** to finish.

Defining Design Variables

If a categorical variable has k values or categories, the system will generate $k-1$ design variables using the Partial method. In the Partial method, each design variable is used with the other design variables to contrast a category with the first category (*i.e.* the reference group).

Variables

The Variables list box lists all currently defined nominal and ordinal variables.

Reference Group

The Reference Group list box lists the category names/values of the variable that is highlighted in the Variables list. Selecting a category from this list controls which category will be used as the reference group. The system creates default design variables that you can either accept or change.

Design Variables

Double clicking on a particular design variable causes a Variable Attributes window to be displayed with Scientific Notation, Field Width and Decimal Place information for the variable.

Names

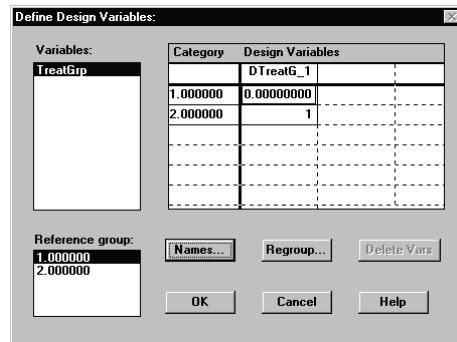
Click on the Names button to display the Design Variables Names window. In that box, you can enter a new name or accept the system default. You can regroup variables into different categories also.

Regroup

Selecting Regroup causes the system to display the Categories window which you can use to regroup variables into different categories. The current category of the variable is displayed in Category. See Modify Categories earlier in this chapter.

Delete Vars

This option is accessible only after you have created design variables in the datasheet. Selecting Delete Vars will delete the design variables associated with the current variable in the datasheet.



Choose the **Variables** menu **Define Variables** option and then the **Design Variables...** option from the datasheet. The Define Design Variables window is displayed.

Link Manager

All of your results (analyses and plots) are linked to the datasheet (or frequency table) from which they were derived. Unless you break the link, changes to the datasheet will be reflected in all of the results derived from it. Similarly, if you make a change to a result in an output window (regression or plot for example), the datasheet being used will reflect the change, and if the change is global, any other connected (linked) results will be affected.

For example, suppose you add a case to a datasheet from which you have obtained three scatterplots, a multiple regression, and descriptive statistics. A new point will appear in the scatterplots, and both the regression and the statistics will be recomputed.

If you omit a point (globally) in one of the scatterplots, the case will disappear from all three scatterplots, and the case number will be grayed out in the datasheet (indicating that it is not in use). Finally, the regression and descriptive statistics will be recomputed without the omitted point.

Your point selection is also reflected in all linked windows. If you highlight a set of points in one of the plots, the corresponding cases will be highlighted in the other plots and in the datasheet.

How the Link Manager Operates

The Link Manager comprises a powerful set of tools that you can use for screening data. Using the facilities provided by the Link Manager, you quickly demonstrate the effects of transforming your data, removing outliers, including or excluding variables, etc. Also, you can readily pick out related cases in various representations of the data. However, you need to understand how the Link Manager operates in order to use it to its full advantage.

The Link Manager maintains linkage between a datasheet, or frequency table, and all of its results, whether they are open or unopened. Certain kinds of changes automatically update all open linked windows immediately. These include adding or deleting cases, changing data values (including redefining a variable using a transformation), and changing a variable name.

Unless the change will invalidate open results, you will receive no warning about these updates. You may want to unlink certain results so that you can compare the results obtained before, and after a modification. Alternatively, you may prefer to copy the original datasheet and maintain two trees in your exploration of the data.

Whether or not you unlink results, it is good practice to save a copy of your original data under a different name until you are sure that you no longer need it. Examples of using the Link Manager are given in Chapter 8 – Tutorial.

Making Local and Global Changes

In some of the plots and analyses you can omit cases locally, as well as globally. A local change is reflected only in the output results in which you make the change. For example; you can locally omit points in the scatterplot and simple regression output results windows.

For a local change, the system re-computes displayed reference lines and statistics for these output results and grays out the point, but no other associated results are affected. This feature allows you to compare results based on alternative selections from the data without unlinking any output results.

A global change is reflected in all open output results. The system re-computes all statistics for all open output results associated with the datasheet.

When Linked Results are Invalid

Some changes may invalidate linked results. For example, deleting or omitting a variable used in a linked analysis may leave nothing to analyze. Changing the number of groups defined by a grouping variable may make an analysis inapplicable, e.g. a *t*-test, which can have only two groups.

Even deleting cases or changing values can sometimes invalidate analyses, e.g. if all values of a variable are equal, or there are no longer enough usable cases for the analysis. When you make a change that will invalidate any linked result, the system prompts you to delete or unlink that result.

Cutting and Pasting Variables and Cases

You may want to perform a cut to reorder or delete variables or cases. For example; you can cut a variable and then paste it next to a related variable so that you can view them together easily. Such operations have no effect on results, as long as you paste the cut elements back into the same datasheet. However, you can also cut an element to delete it, or to move it to another datasheet, and either of those operations could invalidate linked results.

If you cut a case or variable that is essential to maintaining a linked result, the system asks you if you intend to paste the element back into the same datasheet. If you indicate that you do not intend to paste back, the cut will be treated like a delete, and the system prompts you to delete or unlink the invalid result.

If you indicate that you do intend to paste the element back into the datasheet, the system prevents you from doing anything else until you complete the paste operation.

Opening Linked Results

The datasheet and frequency table editor **File** Menus both provide an option to **Open Linked Results**. At any time during a session you can use this option to open saved results. You can open a datasheet and ask to see a linked plot. If the datasheet is changed, it should be noted that all results become invalid and the **Open Linked Results** and **Delete Linked Results** menu options are grayed out.

All datasheet and frequency table windows in the system have the following linkage choices available:

- ◆ Open Linked Results
- ◆ Delete Linked Results

All the result windows in the system have the following linkage choice available:

- ◆ Unlink Results

Open Linked Results

1. Click on **Open Linked Results...** in the **File** Menu. The Open Linked Results window is displayed.
2. Click on the name(s) of the results that you want to open. Each choice is highlighted to indicate your selection.
3. When you are satisfied with your choices, click on **Open Selection**. The system opens all of the result windows that have been selected.
4. Click once to select, click again to deselect.

Unlink Results

1. Click on **Unlink...** in the **File** Menu. The system displays the Name for New Datasheet window.
2. Enter a name for your new datasheet.
3. Click **OK**. At this point, you have not yet saved the result. If you want to save the result, select the **Save Result** option from the **File** Menu in the result window.

Delete linked results

1. Click on the **Delete Linked Results...** option in the **File** Menu. The system displays the Delete Linked Results window.
2. Click on the files you want to delete, then click on **Delete Selection** or **Delete All** as required.

Saving Linked Results

When you close a datasheet, the system asks you if you want to save the results currently open. You can selectively save one or more of these results:

1. To Save All Results: click on the **Save All button**. All results are saved and stored with the datasheet from which they were derived.
2. To Save Selected Results: highlight the results you want to save, then click on **the Save Selection button**. The selected results are saved and stored with the datasheet from which they were derived.
3. To Quit without Saving Any Results: click on the **Do Not Save** button.

Dealing with Open Invalid Results

1. If you make a change that will invalidate currently open linked results, the Unlink Results window is displayed. You can unlink, save, or delete these results.

Unlink and Save Selected Results

1. Drag and drop the result(s) choice(s) from the **These Results are no longer valid** field to the **Results to Save in New File** field. The system gives a default name (FILE1) to the file for the new datasheet that will be linked to the results you are unlinking. This datasheet will have only the variables that are a part of your results.
2. Accept the default name or enter a new name.
3. Click **OK** when you are finished with this window. The **OK** button is enabled when the **These Results are no longer valid** field is empty.

Delete Selected Results

1. Drag and drop the result(s) choice(s) to the **Results to Delete** field.
2. Click **OK** when you are finished with this window. The **OK** button is enabled when the **These Results are no longer valid** field is empty. The results are deleted.

Delete all Results

1. Click on **Delete All**.
2. Click **OK**.

Results made Invalid by Cutting a Variable

If when cutting a variable or case, you invalidate the open linked results:

1. The system displays a message window informing you that the results will be not be valid unless the contents of the clipboard are pasted back into the datasheet.
2. Click on **Yes** to enable you to paste the variable back into the datasheet.

NOTE: If you click on the Yes button, the system will not allow you to do anything until after you have pasted the deleted element back into the datasheet.

3. Click on **No** to delete, or unlink the results.

NOTE: If you click on the No button, the Unlink Results window is displayed.

Unlink and Save all Results

1. Click on **Save All**.
2. All results names appear in the **Results to Save in New File** field, and a default name (FILE1) is given to the file for the new datasheet that will be linked to the results you are unlinking. This datasheet will have only the variables that contribute to your results.
3. Accept the default name or enter a new name.
4. Click **OK** when you are finished with this window.

Results to Save in New File

All results names appear in the **Results to Save in New File** field, and a default name (FILE1) is given to the file for the new datasheet that will be linked to the results you are unlinking. This datasheet will have only the variables that contribute to your results.

Delete all Files

5. Click on the **Delete Linked Results...** option in the **File** Menu. The system displays the Delete Linked Results window, then click on **Delete All**.

Name for New File

The system gives a default name (FILE1) to the file for the new datasheet that will be linked to the results you are unlinking.

Dealing with Unopened Linked Results

At some time, you might have a datasheet with a linked result(s) window that is not currently open. If you make a change to that datasheet, all linked results become invalid, and the **Open Linked Results** option is grayed out in the **File** Menu. This means that you can no longer open those results from the current datasheet.

If you want to preserve the original version of the unchanged datasheet along with its linked results, you should save the changed datasheet with a new name by using the **Save As...** option in the **File** Menu. That procedure ensures that your original datasheet and its linked results remain untouched by your changes.

2. Descriptive Statistics

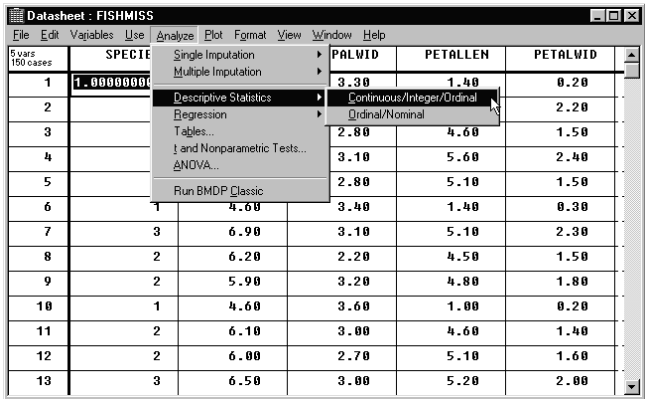
- BASIC SET GROUP
- INDEPENDENCE GROUP
- OUTLIERS GROUP
- NORMALITY GROUP
- ROBUST STATISTICS GROUP

Introduction

The system includes functions for describing distributions of Continuous/Ordinal and Ordinal/Nominal variables.

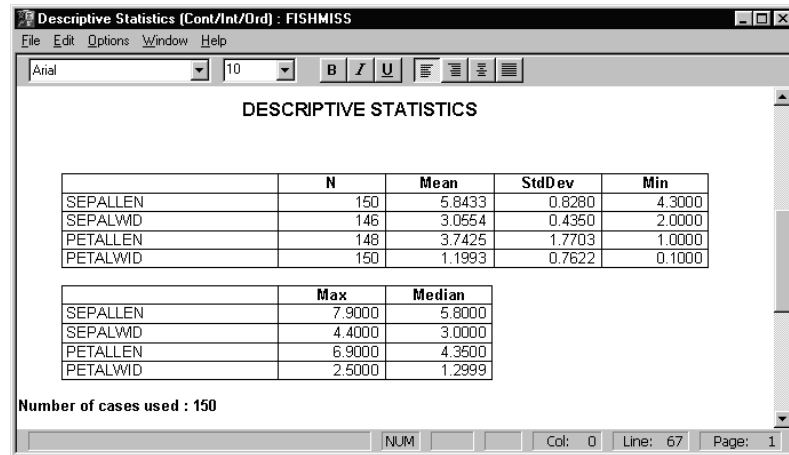
For Ordinal/Nominal variables, the system gives the sample size N, and the proportion of cases for each distinct numerical value. For Continuous/Ordinal variables, five groups of output options are available. Each output option group is described in the following chapter.

From a datasheet menu-bar, select **Analyze**, select **Descriptive Statistics** from the displayed menu, then select **Continuous/Ordinal** from the submenu shown below:



Selecting **Continuous/Ordinal** displays the window shown below:

NOTE: The contents of the windows shown here are displaying data from the system "Samples" folder "FISHMISS.MDD" datasheet.

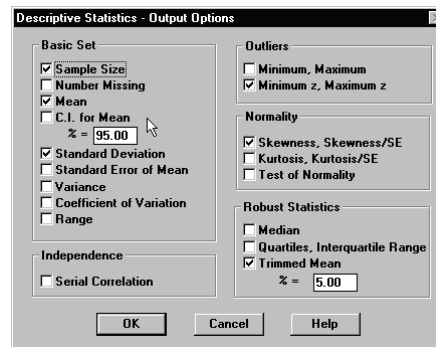


	N	Mean	StdDev	Min
SEPALLEN	150	5.8433	0.8280	4.3000
SEPALWD	146	3.0554	0.4350	2.0000
PETALLEN	148	3.7425	1.7703	1.0000
PETALWD	150	1.1993	0.7622	0.1000

	Max	Median
SEPALLEN	7.9000	5.8000
SEPALWD	4.4000	3.0000
PETALLEN	6.9000	4.3500
PETALWD	2.5000	1.2999

Number of cases used : 150

From the **Descriptive Statistics [Cont/Ord]** window menu-bar select **Options**, then select **Output** from the submenu to display the available options in the **Descriptive Statistics - Output Options** window shown below:



Definitions

The following definitions will be used in the discussion:

N is the effective sample size, or the number of non-missing and valid values of any arbitrary continuous or ordinal variable, called X ;

x_i is the i th value of X , $i=1, \dots, N$;

\bar{x} is the sample mean of X ;

s^2 is the sample variance of X and s is the sample standard deviation;

w_i is an assigned weight to the i th observation, usually equal to 1; and

$t_{(1-\alpha/2, df)}$ is the $(1-\alpha/2)100$ percentile of a Student's t distribution with df degrees of freedom and $0 < \alpha < 1$. That is, the density area to the right of $t_{(1-\alpha/2, df)}$ is $\alpha/2$.

Basic Set Group

The Basic Set Group comprises nine selections each having a check-box. When a function is "checked" (selected), its results will be included in the Descriptive Statistics output window. The **C.I. for Mean** selection has a datafield containing a default value of 95% that can be modified. The selections are described below:

Sample Size

Sample size, N , is the number of cases in the datasheet that have acceptable values for each used variable. Sample size may be different for different variables within a datasheet, since missing data may occur in different cases for different variables. Sample size is less than or equal to the number of used cases in a datasheet.

Number of cases is important, because the behavior of individual statistical measures and entire procedures respond to the absolute size of the sample and/or the relative size of N for distinct variables.

To cite two examples, range has low efficiency for large samples, and standard error of skewness is $(6/N)^{1/2}$ for large N under the assumption of normality. Other analyses, such as multiple regression, require a reasonably large sample size if numerous variables are used.

A more technical definition associated with survey sampling is not implied in our usage here.

Number Missing

Number missing (labeled as **Miss** in the **Descriptive Statistics** Output window) is the number of invalid or missing values of a variable. A "Cases by Variables" datasheet contains a value for each variable when data are complete. Many statistical analyses (such as multiple regression) use a case only if all the variables being considered for admission into the model have no missing values.

To display values for missing data in the Output window, select the **Options** from the **[Cont/Int/Ord]** output window menu-bar, select **Output** to display the Output Options window, then "check" the **Number missing** checkbox.

Missing data can also graphically represented by selecting the **Missing Data Pattern** option from the **View** menu in a Datasheet. This option is described in the Imputation Manual "Missing Data Pattern".

Mean

Arithmetic or sample mean, \bar{x} , is a measure of the central value or central location for the distribution of the used variable. Often called the average, it is the sum of the data values divided by the number of values contributing to that total where:

$$\bar{x} = \sum_{i=1}^N x_i / N$$

The arithmetic mean in the system is actually computed using a provisional means method, whereby data are iteratively centered in order to maintain maximum numerical accuracy. The provisional mean is derived by:

$$\bar{x}_i = \bar{x}_{i-1} + (x_i - \bar{x}_{i-1})w_i / W_i, i = 1, \dots, N$$

where:

\bar{x}_i is the mean of the first i values of X , $\bar{x}_0 = 0$, $\bar{x} = \bar{x}_N$; and:

W_i is the sum of the weights of the first i cases.

The sample mean is an efficient and unbiased estimator for the population mean; i.e., its expected value is the mean of a random variable. If cases or observations are equally weighted (have equal mass), and are positioned along the variable axis, then the mean is identical with the centroid, the center of mass or gravity of the distribution. The sample mean is sensitive to outliers. However, there are robust alternatives to the sample mean, such as the trimmed mean or the median.

Confidence Interval for the mean

The default option provides 95% confidence limits about the mean. Assuming a random sample from a normal population, you are 95% certain of including the true population mean in your confidence interval. These limits are computed as:

$$\bar{x} \pm t_{(1-\alpha/2, N-1)} s / \sqrt{N}.$$

You can enter other confidence levels in the Output Options window.

Standard Deviation

Standard deviation s is the square root of the variance of each variable. It is a widely used measure of variation or dispersion of the observations, since it is in the scale of the original observations. This contrasts with variance which is in squared units.

The usual formula for the unweighted standard deviation is:

$$s = \left[\sum_{i=1}^N (x_i - \bar{x})^2 / (N-1) \right]^{1/2}$$

The system uses a provisional method to iteratively compute the variance (and hence standard deviation) in a case by case manner. This method maintains maximum computational accuracy. See the definition of Variance.

Standard Error of Mean

Standard error of the mean (SE Mean) is the standard deviation (square root of the variance) of the sampling distribution of the mean, rather than the distribution of the observations. The standard error of the mean is obtained by dividing the sample standard deviation by the square root of the sample size:

$$\text{SE Mean} = s / \sqrt{N}$$

SE Mean is used in constructing confidence intervals on the population mean. See Confidence Interval for the mean.

Variance

Variance is the second moment of a distribution taken about the population mean. It is a commonly used measure for variability or dispersion when considering univariate data distributions. It is in squared units of the data.

The sample variance s^2 , an estimate of the population variance, is the sum of the squared deviations from the mean, divided by $N-1$. The sample variance is defined as:

$$s^2 = \sum_{i=1}^N (x_i - \bar{x})^2 / (N - 1).$$

The system uses a provisional method to iteratively compute the variance in a case by case manner so that maximum computational accuracy is maintained. Here:

$$s_i^2 = s_{i-1}^2 + (x_i - \bar{x}_{i-1}) w_i (x_i - \bar{x}_i), i = 1, \dots, N$$

and:

$$s^2 = s_N^2 / (N - 1)$$

where:

s_i^2 is the sum of squared deviations for the first i cases, and $s_0^2 = 0$.

The sample variance is highly sensitive to outliers.

Coefficient of Variation

Coefficient of variation (labeled CoeffVar) is a ratio of the standard deviation of X to its mean:

$$\text{CoeffVar} = s/\bar{x}.$$

It is a dimensionless quantity that is independent of the scale of X . It can be used to compare dispersion in variables where their units of measurement are dissimilar. In practice, it is often used in presenting information on accuracy of a measuring instrument. In that case, CoeffVar is multiplied by 100 and given as a percentage.

CoeffVar is sensitive to the sample estimate of the population mean, most obviously failing to be useful when the mean approaches zero. It assumes that the variable being measured is a ratio variable with a true zero point. It is most useful when the distribution has a natural mean.

Range

Range is the difference between the largest or maximum value and the smallest or minimum value. The numerical value of range tends to increase with increasing sample size.

Range is a simple measure of dispersion, but it is sensitive to outliers. More precisely, range is sensitive to the two extreme observations. In small samples ($N < 10$) range provides a useful nonparametric measure of dispersion. In other circumstances, restricted percentile range measures, such as interquartile range or 10th - 90th percentile range, are preferred. In larger samples, range has low efficiency.

Independence group

The Independence Group comprises one selection named "Serial Correlation" and a check-box. Serial Correlation is described below:

Serial Correlation

Serial correlation (labeled SerCorr) is the Pearson correlation between successive cases in the sample. If serial correlation is high, there may be a lack of independence among cases in a sample. This can be caused by a measuring device drifting out of calibration over time, or a clustering of cases for other reasons.

Serial correlation is computed by computing the correlation of X_i and X_{i+1} for $i=1, \dots, N-1$.

For large N , an appropriate test for the null hypothesis that a population serial correlation is zero can be made using standard normal tables and the following formula:

$$z = [\text{SerCorr} + 1/(N-1)] / \sqrt{(N-2)/(N-1)^2}.$$

If Serial Correlation is "checked" in the Output Options window, the associated p -value(s) for the data set will be displayed when **Continuous/Ordinal** is selected from the datasheet menu-bar.

Outliers Group

The Outliers Group comprises two selections, each with a check-box. The presence of unusual minimum and maximum values for a given variable can indicate the presence of "Outliers" in the data set. Two functions that can indicate this are described below:

Minimum/maximum

Minimum (Min) is the smallest value, and maximum (Max) is the largest value in a set of observations. Minimum and maximum values are commonly examined, and unusual values are indicative of the presence of outliers.

The system can also produce Min and Max across variables within a data case in addition to Min and Max values within a variable.

Minimum/maximum Z-Scores

Minimum and maximum values in a sample are found for a given variable. These are normalized by subtracting the sample mean, then dividing the difference by the sample standard deviation (labeled as Min_Z and Max_Z, respectively). Thus, they may be compared with, or interpreted as though they are, normal deviate scores, or z-scores, following the usual notation used for standard normal (or Gaussian) random variables.

If a variable is normally distributed, you expect to observe a z-score (in absolute value) to be greater than 2 infrequently (approximately 4.6% of the time) and to be greater than 3 very rarely (about 0.3% of the time).

Normality group

Skewness

Skewness is the third moment about the mean in a population. Skewness measures lack of symmetry of the distribution for each used variable. The expected value of skewness is zero for a symmetric distribution. In particular, the normal distribution has zero skewness. Skewness is defined as:

$$\text{Skewness} = \sum_{i=1}^N (x_i - \bar{x})^3 / (Ns^3)$$

Distributions with long upper tails are termed positively skewed, while distributions with long lower tails are termed negatively skewed. The standard error of skewness (SE) is $(6/N)^{1/2}$ for large N under the assumption of normality. The ratio of skewness to its standard error (SKEW/SE) can be interpreted roughly as a standardized score from a normal distribution. This means that absolute values exceeding 2 are unusual (only true for normal variables.)

A wide variety of statistical techniques assume normality of data, but are robust to non-normality, provided (approximate) symmetry is assured. In data analysis, you can reduce positive skewness by applying a variety of transformations, such as square root or logarithmic transformation. However, resulting distributions may still not be fully symmetric. All of the foregoing demand unimodal data.

The numerical value of skewness is highly affected by outliers.

Kurtosis

Kurtosis measures the long-tailedness or peakedness of the distribution of a random variable relative to the Normal or Gaussian distribution with the same mean and variance. It is a dimensionless quantity that is independent of the scale of measurement.

Kurtosis is a simple function of the fourth moment of the distribution of the used variable:

$$\text{Kurtosis} = \left[\sum_{i=1}^N (x_i - \bar{x})^4 / (Ns^4) \right] - 3$$

The standard error of kurtosis (SE) can be approximated by $(24/N)^{1/2}$ for large N and data taken from a normal distribution. Formulas for standard errors of skewness and kurtosis are from Cramer (1946), p. 357.

The measure is standardized such that the Normal or Gaussian distribution has kurtosis = 0. Long-tailed distributions have positive kurtosis. Conversely, distributions more sharply peaked than a normal distribution have negative kurtosis.

The ratio of Kurtosis to its standard error (labeled KURT/SE) may be considered as a standard normal deviate and used to assess normality. A ratio less than -2 may indicate shorter tails than a normal distribution; a ratio greater than 2 may indicate longer tails than a normal distribution. Valid interpretation of the ratio as a normal deviate requires large sample size and no outliers.

Test of Normality

The Shapiro and Wilk statistic (labeled W_Stat) is used to test the hypothesis that a variable has data values drawn from a normally distributed population with mean and variance equal to those in the sample (see Shapiro and Wilk 1965, and Royston 1982).

This test for normality discriminates well between normality and a broad class of alternative distributional hypotheses. The test is effective with quite moderate sample sizes ($N = 30$). The W statistic is available for sample sizes 3 to 2000. The associated p -value is also displayed.

Robust Statistics group

Median

Median is a measure of central tendency. The median is a value that divides the distribution of the data into two equal fractions. Half the data values exceed (or equal) the median, and half are smaller than (or equal to) the median. For this reason, the median is also referred to the 50th percentile, or second quartile.

The median is used with skewed or otherwise non-normal data, and when there are outliers. For normally distributed data, the median has lower relative efficiency than the mean. In this case, the population standard deviation of the median is $\sigma(\pi/(2N))^{1/2}$, where σ is the population standard deviation.

Median computation requires that the data be ordered. Once sorted, the median value is defined as the $(N+1)/2$ th ordered value. If N is even, the median is the average of the $N/2$ th and $(N+2)/2$ th ordered values.

The median is much less sensitive than the arithmetic mean to outlying or extreme values in the data. It also plays a central role in nonparametric tests such as the sign test.

Quartiles and Interquartile Range

Quartiles are three points on the scale of a variable that divide the distribution into quarters or fourths. These are equivalently the 25th percentile or lower quartile (Q1), the median or 50th percentile (Q2), and the 75th percentile, or upper quartile (Q3).

Quartiles are calculated following median computation. Data are ordered, and location of a quartile is determined by $(N+1)p$, where p is the proportion of data falling below a particular quartile (e.g., $p = 0.25$ for Q1 and $p = 0.75$ for Q3). Linear interpolation is used when $(N+1)p$ is not an integer.

Interquartile range, $Q3 - Q1$, and semi-interquartile range, $(Q3 - Q1)/2$, are robust measures of dispersion when central tendency is measured by the median. Semi-interquartile range is also known as quartile deviation.

Trimmed Mean

Trimmed mean or truncated mean (labeled Trim_Mean) is a simple robust form of the mean. Data are ordered and a given percentage of values are removed from each end of the distribution, thereby reducing the sensitivity of the mean to extreme or outlying observations.

For 100p percent ($0 < p \leq 0.25$) trimming, the trimmed mean is calculated as:

$$\text{Trim_Mean} = \left[\sum_{i=k+1}^{N-k} x_i + (1-\varepsilon)(x_k + x_{N-k+1}) \right] / (N - 2pN)$$

where:

$k = 1 + (\text{largest integer less than or equal to } pN)$, and:

$$\varepsilon = pN - k + 1$$

This percentage level of trimming may be used to index the trimmed mean. The system default level of trimming is 5% and the maximum level is 25%.

If a distribution has outlying extreme values, a suitably trimmed mean will differ markedly from the arithmetic mean and signal the presence of these outliers. The trimming process removes equal amounts of the data from the extremities of the distribution in a symmetric fashion until the required fraction is removed.

Limits for the trimmed mean are the usual arithmetic mean for zero trimming, and the median for maximal (50%) trimming. For extreme departure from normality, trimming outperforms the Winsorized mean estimate. See also Hoaglin, Mosteller, and Tukey (1983).

Confidence intervals on the population mean may be derived using the trimmed mean as an estimate for the mean and the Winsorized standard deviation as an estimate for population standard deviation. This combination of estimates is recommended when the underlying distribution is heavy-tailed or is a mixture of normals (see Yuen and Dixon, 1973).

For a discussion on trimmed and Winsorized means, see Dixon and Massey (1983), pp. 380 - 381.

Winsorized Standard Deviation

Winsorized standard deviation (Winsor_SD) is a robust estimate of the population standard deviation and is displayed when the trimmed mean is requested. Winsor_SD is the square root of the Winsorized variance. Given 100p percent trimming, the Winsorized standard deviation is:

$$\text{Winsor_SD} = \sqrt{s_{w_p}^2 / (N - 2m - 1)}$$

where:

$$s_{w_p}^2 = \sum_{i=k+1}^{N-k} (x_i - \bar{x}_{w_p})^2 + k\{[(1-\varepsilon)x_k + \varepsilon x_{k+1} - \bar{x}_{w_p}]^2 + [(1-\varepsilon)x_{N-k+1} + \varepsilon x_{N-k} - \bar{x}_{w_p}]^2\},$$

$$\bar{x}_{w_p} = \left[\sum_{i=k+1}^{N-k} x_i + k\{(1-\varepsilon)(x_k + x_{N-k+1}) + \varepsilon(x_{k+1} + x_{N-k})\} \right] / N$$

$$k = 1 + (\text{largest integer less than or equal to } pN)$$

m is the smallest integer greater than or equal to pN , and:

$$\varepsilon = pN - k + 1$$

The Winsorized standard deviation may be used with the trimmed mean in constructing confidence intervals of the population mean and comparison of means for non-normal data.

3. Regression

SIMPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION OUTPUT OPTIONS

MULTIPLE REGRESSION

MULTIPLE REGRESSION OUTPUT OPTIONS

Regression - General

Regression analysis is used to study relationships between two variables that can be X and Y or between a set of X variables and one Y variable. The X variable(s) is called the independent or predictor variable(s) and the Y variable the dependent or outcome variable. When there is one X variable, the analysis is simple linear regression, and when there are multiple X variables, it is called multiple linear regression.

A mathematical model that predicts a dependent variable from an independent variable(s) represents the relationship.

Two main reasons for fitting a regression model are:

- Predictive:** the equation relating Y and $X(s)$ can be used to predict a given value of Y for one or more given values of X .
- Descriptive:** The type of relationship and its strength may be the main purpose of regression analysis. You may be interested in which X variables relate to Y and in what manner. You may also be interested in the interrelationships among various X variables.

When there is only one X variable, the system will estimate and graph a line that describes the relationship between X and Y . This is simple to visualize, and so the simple linear regression line is given first as a special case in the output. When we have several X s, a hyperplane must be fitted, and it is difficult to visualize the model fit. Thus, we need other methods of examining regression output. Multiple linear regression has been made into a separate option in the system.

The method used to determine a regression line between $X(s)$ and Y is called the least squares method. The least squares method finds the line (plane) that minimizes the sum of squared vertical deviations from each point to the point on the line (plane) corresponding to that case's X value(s). Because it uses squared differences, points that are distant from the line (plane) have a large influence on the placement of the line (plane) particularly if they are extreme points in X . For this reason, great emphasis has been placed on helping users find so-called outlying points.

NOTE: Even though the program is written for linear regression, it can be used in some cases where the relation between X (s) and Y is not linear. This is because the linearity required is between the coefficients, not the original X (s) and the original Y . You can make transformations, such as log, square root or square, on either the original X (s) or Y variable so that a straight line fits the transformed variable(s).

The simple linear regression option is very useful in finding suitable transformations. The multiple regression option can also be used to do simple linear regression, and it contains many features that are not found in the simple linear regression option.

Definitions

The following definitions will be used throughout this chapter:

X_i	Is the i th predictor variable, $i=1, \dots, p$ where p is the number of predictor variables (for simple linear regression, p is equal to 1, and the subscript is omitted).
X	Represents a single predictor variable in simple linear regression. For multiple linear regression, X represents an $N \times (p+1)$ design matrix. Each column of the matrix contains observed values of each of the predictor values, and the first column may be a column of 1s if a model with intercept is requested.
x_j	Is the j th observation of a single predictor variable X , $j=1, \dots, N$.
\bar{x}	Is the sample mean of X .
Y	Represents a dependent variable or an $N \times 1$ vector of observed values of the dependent variable.
y_j	Is the j th observation of Y .
\bar{y}	Is the sample mean of Y .
β_i	Is the i th population parameter associated with the i th predictor variable, $i=0, \dots, p$ (β_0 is the regression intercept).
β	Represents (without subscripts) the $p \times 1$ vector of regression coefficients.
b_i	Is the least squares estimate of β_i .
b	Represents the least squares estimate of β in vector form.
ϵ	Is an error term or an $N \times 1$ vector of error terms, where the e_j 's are independently and identically distributed with zero mean, and variance equal to σ_ϵ^2 .
\hat{Y}	Is the predictor value of Y .
\hat{y}_j	Is the j th predictor value of y_j .
e	Is an estimate of ϵ and is called a residual or vector of residuals.
e_j	Is the j th residual value ($e_j = y_j - \hat{y}_j$), and s_e^2 is an estimate of σ_ϵ^2 .
s_e^2	Is an estimate of σ_ϵ^2 .

Simple Linear Regression

Simple linear regression fits a straight line between Y , the dependent or outcome variable, and X , the independent or predictor variable. The simple linear model is of the form:

$$\hat{Y} = b_0 + b_1 X$$

where b_0 is the estimated intercept, and b_1 is the estimated slope of the line. The sample regression line is fitted to make inferences to the theoretical model given by:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Data used to fit this model are assumed to come from one of two different sampling plans:

1. Random samples of cases are taken at fixed X values, and Y is then measured. Hence, Y is the only random variable.
2. Random samples of cases are taken, and X and Y are measured on each case. Here both X and Y are random variables.

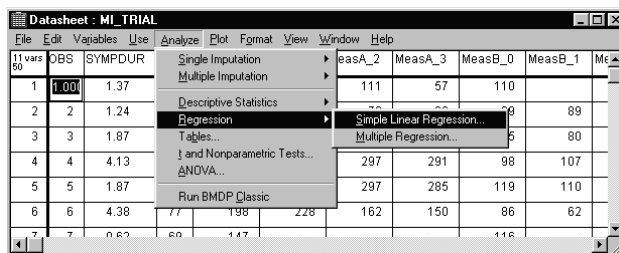
The resulting measurements should be examined in relation to the number of random variables.

It is important to examine the fit of the line to the points to make certain that a straight line is a reasonable model to assume and that there are no obvious outliers. In addition to fitting a straight line, you can perform tests of hypotheses and construct confidence limits. In performing these tests and constructing confidence limits, the residual error terms are assumed to have a normal distribution:

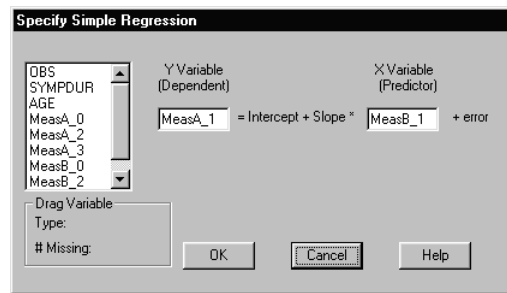
$$(\varepsilon_j \sim N(0, \sigma_\varepsilon^2)).$$

Using Simple Linear Regression

Choose **Simple Linear Regression** from a datasheet **Analyze** menu as shown below:



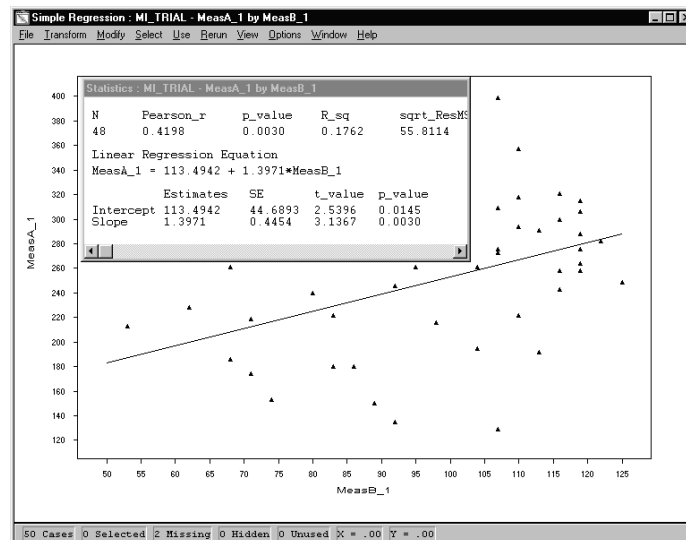
The window shown below is displayed:



To perform Simple Linear Regression:

1. Select **Simple Linear Regression** from a datasheet **Analyze** menu to display the Specify Simple Regression window.
2. Drag the desired variable into the **Y Variable (Dependent)** datafield.
3. Drag the desired X variable into the **X Variable (Predictor)** datafield.
4. Click **OK**.

A Scatterplot with the regression line is displayed in the Simple Regression Output window.



Modifying the Output

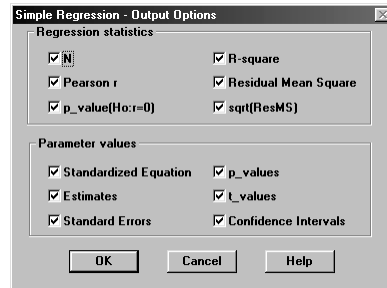
You can modify the Simple Linear Regression output using the menus in the Output window to:

- ◆ Select points for possible deletion using the **Select** menu.
- ◆ Distinguish points by group using the **View** menu.
- ◆ Make transformations using the **Transform** menu.

You might delete points if you consider them to be outliers. Transformations can be considered when there is a poor fit between the points in the Scatterplot and the fitted line.

Simple Linear Regression Output Options

From a Simple Linear Regression Output window (shown above), select **Options** → **Output** to display the Simple Regression – Output Options window:



Clicking a checkbox in the Regression Statistics group, or the Parameter values group, and pressing the **OK** button will include the result for the selection(s) in the Output window as shown below.

Regression Statistics

The Regression Statistics and Parameter value results output window is shown below:

Statistics : MI_TRIAL - MeasA_1 by MeasB_1						
N	Pearson_r	p_value	R_sq	ResMS	sqrt_ResMS	
48	0.4198	0.0030	0.1762	3114.91	55.8114	
Linear Regression Equation						
MeasA_1 = 113.4942 + 1.3971*MeasB_1						
Standardized Equation						
z_score(MeasA_1) = 0.4198 * z_score(MeasB_1)						
	Estimates	SE	t_value	p_value	Lower_CI	Upper_CI
Intercept	113.4942	44.6893	2.5396	0.0145	23.5392	203.4492
Slope	1.3971	0.4454	3.1367	0.0030	0.5006	2.2937

Regression Statistics Formulae

The formulae used for Regression Statistics are described below:

N

N is the sample size or effective number of cases used in fitting a regression line.

Pearson r

Pearson r is the usual product-moment correlation coefficient of Y and X . It is also the correlation between Y and its predicted value \hat{y}_j ;

$$r = \frac{\sum_{j=1}^N (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^N (x_j - \bar{x})^2} \sqrt{\sum_{j=1}^N (y_j - \bar{y})^2}}$$

The test statistic below has a Student's t distribution with $N - 2$ degrees of freedom:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

When you request it, the associated p -value is given for the test of the null hypothesis that the population correlation is zero. This p -value is equivalent to the p -value obtained from an F -test in the ANOVA table. In making statistical tests concerning the population correlation coefficient, you are assuming that a sample of cases was taken from a bivariate normal population. The correlation coefficient has the advantage of being independent of the units of measurement:

p-value

p - value ($H_o: r = 0$) See discussion under Pearson r .

R-square

R-square (labeled as R_sq) is the square of the Pearson r . It is also called the coefficient of determination and may be calculated as: the ratio of the regression sum of squares to the total sum of squares. It measures the proportion of explained variation based on a regression model.

Residual Mean Square

Residual Mean Square (s_e^2) is the sum of squared vertical deviations of the points about the regression line divided by $N - 2$. It is an estimate of the variance of error terms in a regression model:

$$s_e^2 = \sum_{j=1}^N (y_j - \hat{y}_j)^2 / (N - 2).$$

Sqrt_ResMS

The standard error of the estimate (labeled sqrt_ResMS) is the square root of the residual mean square. It is an estimate of the standard deviation of the error terms.

$$\text{Sqrt_ResMS} = s_e = \sqrt{\sum_{j=1}^N (y_j - \hat{y}_j)^2 / (N - 2)}.$$

Parameter values

For each regression coefficient, you can request its estimate, standard error, p -value, t -value, and confidence interval. Results of each selection for Parameter values in the Output Options window will be displayed (together with Regression Statistics) in an output window as shown above.

Parameter values Formulae

Output results for Parameter values are shown above, and the formulae are described below:

Intercept

The sample intercept b_0 is the distance from the origin to the place where the regression line cuts the Y axis. It is derived as:

$$b_0 = \bar{y} - b_1 \bar{x}.$$

The standard error of b_0 (labeled as SE) is the square root of the estimated variance of b_0 :

$$SE(b_0) = \left[s_e^2 \left\{ (1/N) + (\bar{x}^2 / \sum_{j=1}^N (x_j - \bar{x})^2) \right\} \right]^{1/2}.$$

The t -value is defined as:

$$t - \text{value} = b_0 / SE(b_0).$$

and is distributed as a Student's t with $(N - 2)$ degrees of freedom. The t - and p - values are used for the test of the null hypothesis that the population slope β_0 is zero. In testing that β_0 is zero, you are also assuming that the error terms are normally distributed.

The 95% confidence limit for b_0 is defined as:

$$b_0 \pm t_{(.975, N-2)} SE(b_0).$$

Slope

Slope b_1 is the estimated change in Y for one unit of change in X . The size of b_1 will depend on the units of X defined. It is defined as:

$$b_1 = \sum_{j=1}^N (y_j - \bar{y})(x_j - \bar{x}) / \sum_{j=1}^N (x_j - \bar{x})^2.$$

Standard error of b_1 (labeled as SE) is the square root of the estimated variance of b_1 :

$$SE(b_1) = \left[s_e^2 / \sum_{j=1}^N (x_j - \bar{x})^2 \right]^{1/2}.$$

The t -value is defined as:

$$t - \text{value} = b_1 / SE(b_1)$$

and is distributed as a Student's t with $(N - 2)$ degrees of freedom. The t - and p - values are used for the test of the null hypothesis that the population slope β_1 is zero. In testing that β_1 is zero, you are also assuming that the error terms are normally distributed.

The 95% confidence limit for b_1 is defined as:

$$b_1 \pm t_{(.975, N-2)} SE(b_1)$$

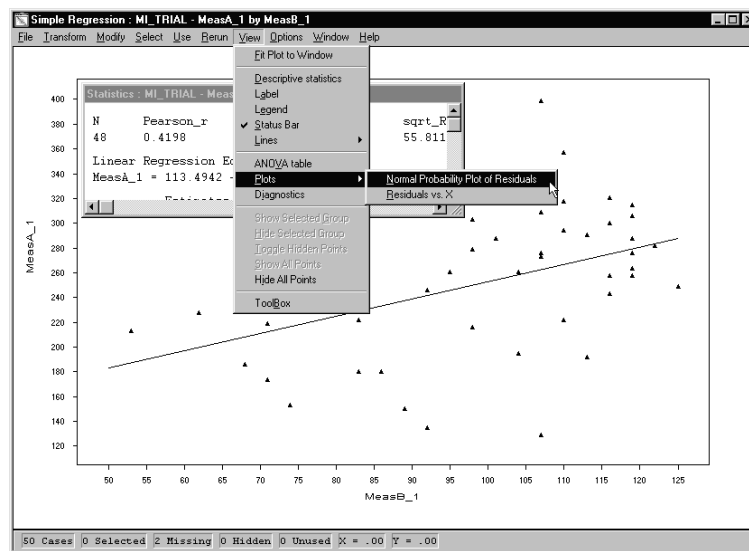
Standardized Regression Equation

Standardized regression coefficients are regression coefficients obtained from data that are standardized such that all means equal zero and all variances equal one. This standardization can be accomplished by subtracting the sample mean from each observation and dividing the difference by the standard deviation. For simple linear regression, the standardized slope coefficient is equal to r and the intercept is zero. The standardized regression equation is the equation with the standardized coefficients.

Additional Output – View menu

Selecting the **View** menu in an Output window displays four additional output options:

- ◆ Descriptive Statistics
- ◆ ANOVA
- ◆ Plots
- ◆ Diagnostics



Descriptive Statistics

Output includes the sample size, the mean, standard deviation, minimum, maximum, median, first and third quartiles of the X and Y variable. Refer to Descriptive Statistics in Chapter 2. The Descriptive Statistics output window is shown below:

Descriptive Statistics

	N	Mean	StdDev	Min
MeasA_1	48	251.3750	60.8334	129.0000
MeasB_1	48	98.6875	18.2770	53.0000

	Max	Median
MeasA_1	399.0000	261.0000
MeasB_1	125.0000	104.0000

Number of cases used : 48

ANOVA Table

An analysis of variance (ANOVA) table decomposes the total sum of squares into two sums of squares. The first sum of squares is the regression sum of squares (the explained part), and the second is the residual sum of squares (the unexplained portion). Each of these two sources of variation has associated degrees of freedom. Formulas for the ANOVA table are summarized below:

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	$\text{RegSS} = \sum_{j=1}^N (\hat{y}_j - \bar{y})^2$ $= b_1 \sum_{j=1}^N (y_j - \bar{y})(x_j - \bar{x})$	1	$\text{RegMS} = \text{RegSS}$	$F = \frac{\text{RegMS}}{\text{ResMS}}$
Residual	$\text{ResSS} = \sum_{j=1}^N (y_j - \hat{y}_j)^2$	$N - 2$	$\text{ResMS} = \frac{\text{ResSS}}{N - 2}$	

The null hypothesis tested in this ANOVA table is the hypothesis of zero population slope ($H_0: \beta_1=0$ vs. $H_1: \beta_1 \neq 0$.) To construct an F test, a ratio of the Regression Mean Square (RegMS) to the Residual Mean Square (ResMS) is derived. The test statistic F -value has an F distribution with 1 and $(N-2)$ degrees of freedom.

An additional column (labelled p _value) shown in the system output provides the associated p -value. The p -value is the density area to the right of the test statistic value, based on an F distribution with 1 and $(N-2)$ degrees of freedom. If the p -value is less than a predetermined level of significance (e.g., 0.05), then the null hypothesis is rejected. The p -value assumes that the null hypothesis is true.

An ANOVA results output window is shown below:

Analysis of Variance

	Sum of Squares	DF	Mean Square	F value	p value
Regression	30647.2700	1	30647.2700	9.8389	0.0030
Residual	143286.0000	46	3114.9130		

Diagnostics

The diagnostics datasheet contains the X and Y values, predicted values, $\hat{Y}(= b_0 + b_1 X)$, residuals $e(= Y - \hat{Y})$, and prediction intervals.

A 100(1- α) percent prediction interval for the j th observation is computed as:

$$\hat{y}_j \pm t_{(1-\alpha/2, N-2)} s_e \left[1 + (1/N) + \{(x_j - \bar{x})^2 / \sum_{j=1}^N (x_j - \bar{x})^2\} \right]^{1/2}$$

where $t_{(1-\alpha/2, N-2)}$ is the 100(1- $\alpha/2$)th percentile of a Student's t distribution with $(N-2)$ degrees of freedom. This interval is a prediction interval of a single value of Y at X , and should not be confused with the confidence interval of the expected value of Y given X .

An Output window showing the Diagnostics results is shown below:

Case	MeasA_1 YVar	Predicted	Residual	PI_Lower	PI_Upper	MeasB_1 XVar
1	150	237.84	-87.84	124.00	351.68	89
2	240	225.27	14.73	110.53	340.00	80
3	276	262.99	13.01	149.24	376.74	107
4	294	267.18	26.82	153.22	381.14	110
5	228	200.12	27.88	81.94	318.29	62
6	321	275.56	45.44	161.00	390.13	116
7	213	187.54	25.46	66.87	308.21	53
8	216	250.41	-34.41	136.91	363.92	98
n	200	270.75	0.00	144.00	396.74	110

NOTE: You can request a confidence band for the regression line from the **View** menu **Lines** option.

Plots for Simple Linear Regression

The **Plots** option includes a **Normal Probability Plot of Residuals** and a plot of **Residuals vs X**. You can also use the plot of the residuals against X as a check for outliers and for the fit of the regression line.

You can use the normal probability plot to show graphically whether the residuals appear to be normally distributed. Residuals are plotted on the horizontal axis, and the corresponding expected normal values (based on ranks) plotted on the vertical axis. If the residuals are almost normally distributed, the sample points will lie approximately on a straight line.

Multiple Regression

Multiple linear regression fits a linear model using one or more predictor variables, X , to predict one dependent variable, Y . A regression model with p X 's is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

where:

the β_i 's are the unknown parameters, and ε is a random component with zero mean and variance σ_ε^2 . Each observation is assumed to be independent of the rest.

The β_i 's ($i=1, \dots, p$) represent the slope parameters of a regression on a hyperplane with respect to the p X s.

The β_i 's are called regression coefficients. For example, β_1 is the rate of change of the mean of Y as a function of X_1 when levels of the remaining X s are held fixed. The parameter β_0 represents the intercept.

NOTE: There are $(p + 1)$ parameters being estimated unless the intercept is fixed at zero.

Data used to fit this model are assumed to come from one of two different sampling plans:

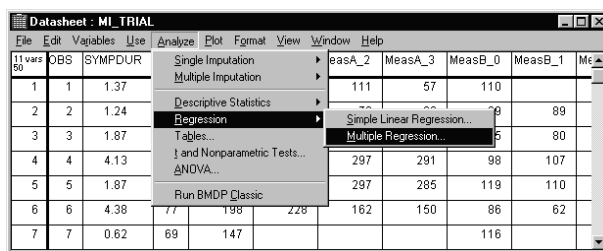
1. Random samples of cases are taken at fixed X values, and Y is measured. Here Y is the only random variable.
2. Random samples of cases are taken, and the X s and Y are measured on each case.

For the first sample plan, the only correlation that is meaningful is the multiple correlation between the dependent variable Y and its predicted value \hat{Y} based on a multiple linear regression.

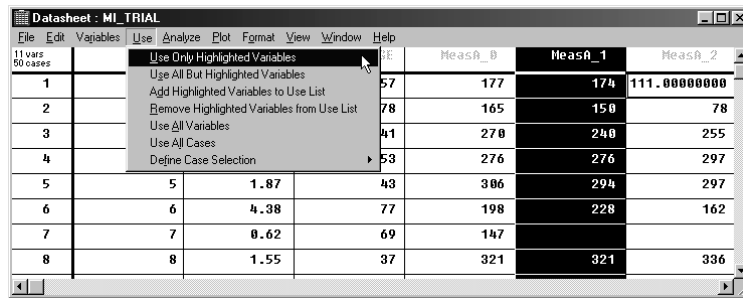
Another equivalent measure is the multiple correlation squared (labeled as R_{sq}), also known as the coefficient of determination. This can be used as an indication of the reduction in the variance of Y due to a regression of Y on the X s.

Using Multiple Regression

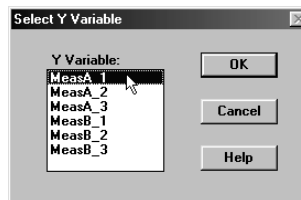
Multiple regression estimates of β_i s are obtained using the method of least squares. The **Multiple Regression** option in the datasheet **Analyze** menu allows you to enter variables in an iterative fashion.



If your datasheet contains a large number of variables, the **Use** menu in the datasheet allows you to select a subset of variables/cases for possible use in the analysis. Pruning the variable list in this way will save you from scrolling through numerous variables that you do not wish to use in your regression analysis.



The **Use** menu allows you to limit the number of variables/cases that are scanned for missing values. Only cases with complete data in the scanned variables will be used in the regression analysis. Thus, it is important to limit your analysis to the minimal set of variables of interest. Selecting **Multiple Regression** from the datasheet **Analyze** menu displays the window shown below:



After you select the *Y* variable and press the **OK** button, the Multiple Regression Output window is displayed:

Multiple Regression : ML_TRIAL - MeasA_1

File Edit Model View Options Run Window Help

N	Multiple_R	R_sq	Adj_R_sq	sqrt_ResMS	
39	0.9125	0.8327	0.8282	22.7467	

Multiple Linear Regression Equation

MeasA_1 = 49.7094 + 0.8155*MeasA_2

Variables in model

Variable	Coeff	SE_coeff	F_to_R	p_value	t_value
MeasA_2	0.82	0.06	184.18	6.147E-16	13.57

Variables not in model

Variable	Partial_r	F_to_E	p_value	t_value	
OBS	-0.0758	0.21	0.6511	0.46	
SYMPDUR	-0.0058	1.23E-03	0.9722	0.04	
AGE	0.3145	3.95	0.0544	1.99	
MeasA_0	0.5992	20.17	7.038E-05	4.49	
MeasA_3	-0.0355	0.05	0.8324	0.21	
MeasB_0	-0.0311	0.03	0.8531	0.19	
MeasB_1	-0.1859	1.29	0.2637	1.14	
MeasB_2	-0.1889	1.33	0.2561	1.15	
MeasB_3	-0.1680	1.05	0.3133	1.02	

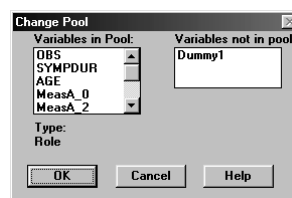
There will be no X variables in the model if all candidate X variables have been assigned roles of None in the Specify Variable Attributes window. If variables are assigned the role of X , they will automatically be included in the model. Here the variable **MeasA_2** has been given the role X in the Specify Variable Attributes window shown below:

Specify Variable Attributes			
Basic Attributes Outpoints Copy Attributes			
Variable Name:	Type:	Role:	
MeasA_2	Continuous	X variable	
Variables:	Format:	Case Frequency Case Label X variable Y variable	
OBS SYMPDUR AGE MeasA_0 MeasA_2 MeasA_3 MeasB_0 MeasB_1 MeasB_2 MeasB_3	Alignment: Decimal Field Width: 9 <input type="checkbox"/> Scientific Notation <input type="checkbox"/> Show Group Names Decimal Places: 0 <input type="checkbox"/> Double Precision	Equation:	
<input type="button" value="New Variable"/> <input type="button" value="OK"/> <input type="button" value="Cancel"/> <input type="button" value="Help"/>			

If you have declared a variable to be nominal, then you need design variables (or dummy variables) in order to use this variable as a predictor variable in Multiple Regression. The **Define Variables → Design Variables** option in the datasheet **Variables** menu allows for this possibility and will create design variables for you. See “Creating Design (Dummy) Variables” in the next section.

You can not, however, use a nominal variable as a dependent variable. If you wish to use a nominal variable as a dependent variable, consider the Classic BMDP logistic regression programs, LR and PR.

If you pre-assign some of the variables as X variables, they will be automatically entered into the regression equation. You may also select or deselect variables under the **Change Pool** option in the **Model** menu of the Output window. If they turn out to be poor predictors, you can remove them by dragging them out of the **Variables in Pool** list in the Change Pool window shown below:



Creating Design (Dummy) Variables

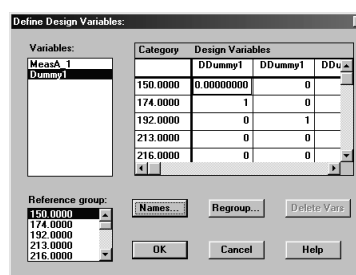
If you have nominal data and wish to use a nominal variable as an X variable, and you have selected the **Multiple Regression** option, the system displays a message window shown below.

The system will allow you to generate design variables from a nominal variable or treat the nominal variable as if it were continuous, or remove the variable from the pool. For an ordinal variable, you must change its type to nominal if you want to use dummy variables. Dummy coding puts less stringent assumptions regarding the relationships between that categorical variable and the dependent variable.

You can create design variables in two ways:

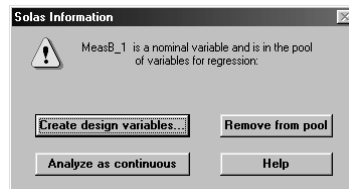
Starting from the Variables menu.

1. Click on the **Variables** menu **Define Variables** option. Then choose the **Design Variables...** option. The system displays the Define Design Variables window containing default design variables. The dialog box displays one less design variable than the number of categories for your nominal variable.
2. Click on **OK** when you are satisfied with your coding choice. A set of non-redundant design variables is created. It is possible to change the grouping using the **Regroup** option and use the **Names...** option to change the names of the design variable.



Starting from the Analyze menu Multiple Regression option.

1. When you choose the **Analyze** menu **Multiple Regression** option, and your datasheet includes one or more nominal variables, you will see the SOLAS™ information message window shown below. Click on the **Create Design Variables...** option, and the system displays the Define Design Variables window (shown above) containing default design variables. The dialog box displays one less design variable than the number of categories for your nominal variable.
2. Click on **OK** when you are satisfied with your coding choice. A set of non-redundant design variables is created. You can change grouping using the **Regroup** option and use the **Names...** option to change the names of the design variable.



Design Variables can be coded in several different ways:

If a nominal variable has K levels, then you will get $K-1$ dummy variables. Thus, the use of numerous nominal variables in Multiple Regression will lead to many variables in the model.

Suppose patients are classified as normal, disease A, or disease B, and you decide to use normal as the reference group. The assigned data for these three groups are as follows:

Groups	D1	D2
reference-normal	0	0
disease A	1	0
disease B	0	1

There are $K-1$ or 2 dummy variables created. The first dummy variable takes on a value of 1 if the case has disease A, and zero otherwise. The second dummy variable takes on a value of 1 if the case has disease B, and zero otherwise.

The estimate of the coefficient for D1 is sometimes called the differential intercept coefficient since it tells how much the intercept term for patients with disease A differs from the intercept term for patients classified as normal (the reference group). Likewise, the coefficient for D2 tells how much the intercept term for patients with disease B differs from that for patients classified as normal.

NOTE: This is the same as the partial method given in the Classic BMDP logistic regression program, LR. If dummy coding, you must assign one of the levels to be the reference group. You should choose the reference group such that the interpretation of the coefficients is most meaningful. Also, it should have a sufficient number of cases, because all the other groups will be compared to it (see Kirk, 1982, pp. 186-187 or Afifi and Clark, 1990, pp. 225- 133).

Multiple Regression Output Options

You can modify your options for the Regression output from the **View** menu **System Preferences** option in the SOLAS™ Main window. A variety of output options are available for Multiple Regression and are described in the following sections.

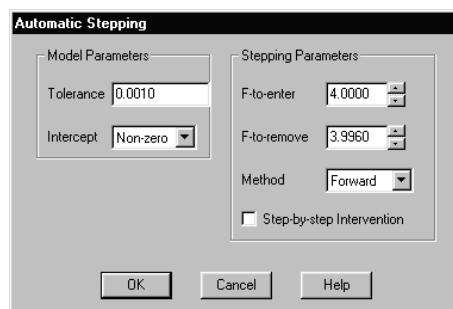
Model Menu

When the Multiple Regression Output is displayed, a variety of options are available from the **Model** menu where you can change the Model Parameters, enter or remove variables from the Regression Pool etc.

Modifying the Output

There are several methods for entering X variables. One widely used choice is forward stepwise which you do by clicking on the **Model** menu in the Output window and then choosing **Automatic Stepping**.

The automatic stepping defaults include: tolerance set at 0.001, intercept chosen to be non-zero, the F -to-enter value set at a default value of 4.0, and method is chosen to be forward (*i.e.*, forward stepwise). If you wish to see more variables entered in the model, set the F -to-enter value to a smaller value. The numerical value of F -to-remove should be chosen to be less than the F -to-enter value.



Alternatively, you can enter variables yourself in an interactive fashion. You do this directly from the Multiple Regression Output window.

Variable	Partial_r	F_to_E	p_value	t_value
OBS	-0.0687	0.17	0.6861	0.41
SYMPDUR	-0.0144	7.24E-03	0.9327	0.09
AGE	0.3719	5.62	0.0234	2.37
MeasA_0	0.6051	20.22	7.258E-05	4.50
MeasA_3	-0.0282	0.03	0.8684	0.17
MeasB_0	0.2866	3.13	0.0855	1.77
MeasB_2	-0.0580	0.12	0.7333	0.34
MeasB_3	-0.0392	0.05	0.8177	0.23

In the Multiple Regression Output window, the Variables not in Model window shown above contains candidate variables that you can enter into the model. If you are certain about the variables that you want to use in the regression, drag these variables from this window and place

them into the Variables in model window shown below. Also, if you have prior justification for using particular variables, but are uncertain about other variables, place the variables about which you are certain in the Variables in model window.

Variables in model					
Variable	Coeff	SE_coeff	F_to_R	p_value	t_value
MeasR_2	0.86	0.07	148.07	2.581E-14	12.17
MeasB_1	-0.28	0.25	1.29	0.2637	-1.14

After you enter specific variables in the model, or if you do not have any such variables, it is time to inspect the Variables not in model window. There you find information on the variables that have high partial correlation r (labeled as Partial_r) in absolute value or large F -to-enter values (labeled as F _to_E). One common practice is to enter next the variable with the highest F -to-enter value.

If none of the variables has a sizable F -to-enter value you may not wish to enter any of them. A commonly used cut-off point for F -to-enter values is an F -value equivalent of a p -value of 0.10 to 0.25 when the purpose of fitting a regression model is predictive. Values to use in practice are a minimum F -to-enter value of 2.07 (see Bendel and Afifi, 1977, pp. 46 - 53) or, for a smaller p -value, the default F -value of 4.0.

You can continue entering more variables, one at a time, choosing those with the highest F -to-enter values, or you can mix this method of entry with the method of entering those variables you prefer for conceptual reasons. The process stops when none of the remaining X variables has a large enough F -to-enter value, or you run out of variables. Regression models built in this manner are called explanatory models. The p -value associated with an F -value should be interpreted cautiously since multiple tests have been made.

As you enter variables into the multiple regression model, the regression coefficients will change as will the multiple correlation, its square, and the adjusted multiple correlation. The system updates the multiple regression equation as you move variables in and out of the model.

NOTE: The estimated regression coefficients are printed in front of the variable to which they apply in the equation.

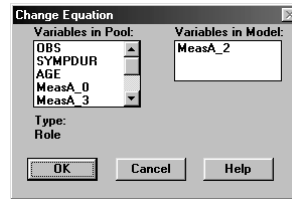
Model Menu

The **Model** menu provides ways to:

- Change the equation by entering and removing X variables.
- Change the regression model parameters.
- Change the automatic stepping parameters.
- Change the variables in the pool.

Change Equation

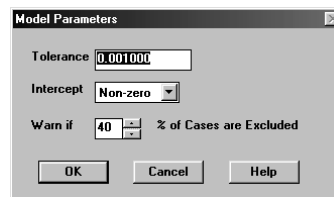
Selecting **Change Equation** from the **Model** menu displays the window shown below. This option provides another way of entering and removing X variables in the model.



Change Model Parameters

Selecting **Change Model Parameters** from the **Model** menu displays the window shown below. This option allows you to:

- Set a minimum allowed tolerance before a warning is issued.
- Force a zero intercept or allow the intercept parameter to be estimated for the sample.
- Specify the percent of cases which causes a warning to appear for the cases being excluded for missing values.



Tolerance

The nature of the correlation structure among the X variables is measured by tolerance, a crucial factor in assessing stability of regression estimates. Tolerance is defined as 1 minus the squared multiple correlation between any X variable and the remaining X variables in the model. For a given X variable, the system computes the tolerance between that variable and all currently entered variables. The program also computes the tolerance between each currently entered variable with all variables in the model as the tolerance would be if the current candidate variable were included. Variables that enter into the model must have tolerance values greater than the tolerance limit set under Change Model Parameters.

Tolerance is used as a measure of multicollinearity. When multicollinearity is present, computed values of the regression coefficients may be inaccurate and unstable. Small values of tolerance, say less than 0.001, may indicate problems.

Intercept

Intercept is defined as the distance from the origin to the place where the regression hyperplane cuts the vertical axis. This option allows you to set the intercept at zero or have the program compute the intersection of the least squares regression hyperplane with the vertical axis. The intercept may also be moved in or out of the model.

A zero-intercept model is used when you hypothesize that the regression line should pass through the origin.

When you select the zero-intercept model, uncorrected sums of square are used in calculation.

Warn if [] % of cases excluded

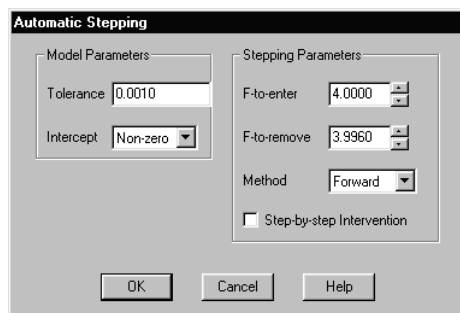
This option allows you to change the percent (0 - 90) which, when exceeded, causes the system to warn you that the percent of excluded cases has exceeded your set percent. Cases will be excluded if there are any missing values in the pool of variables (dependent and independent) being used.

They may also be excluded if you set up special conditions for exclusion. The default is 40%.

Any regression model being computed will use only cases with complete data for the pool of variables used. When there is a high proportion of excluded cases, it becomes questionable to make inferences from the sample to a population. You can use the missing data matrix to see if a sizable proportion of the missing values occur in one or two variables. When missing values occur in only a few variables, you might consider removing these variable(s) from the pool by utilizing the Use Menu from the datasheet.

Automatic Stepping

Using the **Automatic Stepping** option in the Multiple Regression output window **Model** menu you can set the Model and Stepping Parameters.



For definitions of *F*-to-Enter and *F*-to-Remove see the section “Formulae” later in this chapter.

For definitions of Tolerance and Intercept see the section “Model Parameters” above.

Step by step Intervention

Variables with the highest *F*-to-enter values will be entered one at a time. You can control whether or not the next step proceeds.

Method

In the forward stepwise method, a step consists of adding a variable to the regression model. Forward stepwise variables are automatically added one at a time into the model (and possibly some variables in the model may be removed) dependent on your stated criteria until a stopping point is reached.

When entering variables into a regression model, you use *F*-to-enter as your criterion. You add the variable with the largest computed *F*-to-enter value at that step. Variables already in the model are also checked, and if any of them have an *F*-to-remove value less than the stated *F*-to-remove value they will be removed. The procedure continues until no variables have either an *F*-to-enter value larger than the stated *F*-to-enter value in this window or an *F*-to-remove value less than the stated *F*-to-remove value.

The F -to-enter value must be chosen to be larger than the F -to-remove value (see Afifi and Clark, 1990, pp. 196 - 203 or any of the numerous other regression texts).

NOTE: Forward stepwise regression is not guaranteed to result in the best possible regression equation. In particular, when the number of cases is small relative to the number of variables, spurious variables may enter.

Change Pool

See “Using Multiple Regression” above.

View Menu

Variables Not in Model

Highlight the selection in the **View** menu to display the window.

Descriptive Statistics Window

Output is included from the descriptive statistics model/object. Output includes the sample size, the mean, standard deviation, minimum, maximum, median, first and third quartiles of the X and Y variable. Refer to Descriptive Statistics in Chapter 2. The Descriptive Statistics output window is shown below:

	N	Mean	StdDev	Min
OBS	39	26.8461	14.8794	2.0000
SYMPDUR	39	2.5553	2.1798	0.2800
AGE	39	53.9487	13.9368	24.0000
MeasA_0	39	254.9230	49.9227	153.0000
MeasA_1	39	247.6923	54.6779	135.0000
MeasA_2	39	242.7692	61.4060	78.0000
MeasA_3	39	237.0000	68.2684	33.0000
MeasB_0	39	99.6923	16.8275	62.0000
MeasB_1	39	97.5384	17.4051	62.0000
MeasB_2	39	93.2307	18.3997	56.0000
MeasB_3	39	89.5384	20.0287	44.0000

	Max	Median
OBS	50.0000	26.0000
SYMPDUR	12.1599	1.9600
AGE	78.0000	56.0000
MeasA_0	321.0000	270.0000
MeasA_1	359.0000	258.0000
MeasA_2	354.0000	252.0000
MeasA_3	345.0000	252.0000
MeasB_0	125.0000	101.0000
MeasB_1	125.0000	98.0000
MeasB_2	122.0000	95.0000
MeasB_3	125.0000	89.0000

Number of cases used : 39

ANOVA table

The ANOVA table is provided for testing the hypothesis that all regression coefficients are zero. If variables are entered based on a conceptual model and not using information from the sample in a stepwise or other fashion, the F - and p -values will have their regular interpretation (assuming the usual assumptions are met).

If variables are entered in a stepwise fashion using information from the sample to choose variables that maximize the F -to-Enter value or some other criterion at each step, then the p -

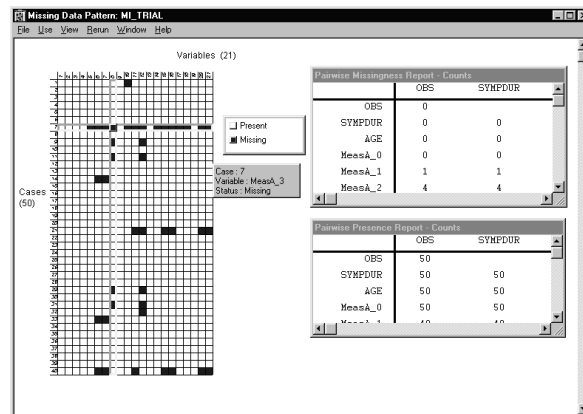
value associated with the F statistic cannot be trusted (see Forsythe, A.B. *et al.*, 1973). Formulas for the ANOVA table are summarized as follows:

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	$\text{RegSS} = \sum_{j=1}^N (\hat{y}_j - \bar{y})^2$	p	$\text{RegMS} = \frac{\text{RegSS}}{p}$	$F = \frac{\text{RegMS}}{\text{ResMS}}$
Residual	$\text{ResSS} = \sum_{j=1}^N (y_j - \hat{y}_j)^2$	N-p-1	$\text{ResMS} = \frac{\text{ResSS}}{N - p - 1}$	

The F -test is the regression mean square (RegMS) divided by the residual mean square (ResMS). A large F -value indicates that the values of Y fit closely to the regression model even though the precise p -values cannot be trusted if variables are entered based on simple statistics.

Missing Data Pattern

Selecting this option from the **View** menu in a datasheet allows you to determine, at a glance, which variables have missing values before you perform the Multiple Regression. From the **Use** menu, you can select/de-select variables and cases for the analysis.



Corr/Cov Matrices

Correlation of covariance matrices is given for variables in the pool and for regression coefficients.

Correlation Matrix of Variables

The Correlation Matrix of Variables in the pool displays a correlation matrix with the dependent variable in the first row and column. Correlations given here are the usual product moment correlations. The correlation between any two variables can be found by finding one of the variables in a row and the second one in a column and then finding the desired correlation by proceeding right in the chosen row until you intersect the chosen column. The matrix is derived from the covariance matrix of variables.

When the numerical value of the tolerance is small, or there are other indications of multicollinearity, it is often useful to examine the correlation matrix to see which variable(s) are causing the problem.

Covariance Matrix of Variables

Covariance Matrix of Variables displays a covariance matrix with the dependent variable in the first row and column. The sample variances of the variables are given in the diagonal and the covariance in the off diagonal.

The covariance matrix is sometimes saved when there is a large sample size. It can be used in subsequent regression analyses to save computation time.

When the zero -intercept model is used, the covariance matrix is uncorrected for the mean. For example, the variance of Y is:

$$\text{Var}(Y) = \sum_{i=1}^N Y_i^2 / N.$$

Correlation Matrix of Coefficients

Correlation matrix of the estimated regression coefficients is displayed. This can be used to determine whether any of the slope coefficients are highly correlated. This is derived from the covariance matrix of coefficients.

Covariance Matrix of Coefficients

Covariance matrix of the estimated regression coefficients is displayed. This can be used to compute confidence limits and test hypotheses of chosen slope coefficients. The covariance matrix is calculated as $\text{cov}(b) = \text{ResMS}(X'X)^{-1}$ where ResMS is the residual mean square.

Plots

From the Regression Output window **View → Plots** menu you can select three plots:

- ◆ Diagnostic Plots
- ◆ Partial Plots
- ◆ Custom Plots

Diagnostic Plots

Selecting the Diagnostic Plots option results in a plot of the residuals versus the predicted values. This is a general-purpose plot that is often examined for outliers, lack of model fit, or lack of normality of error terms.

Partial Plots

Two Partial Plots are available. These plots allow you to study the effects of an additional variable in a regression model. They are often used to evaluate variables being added to the model in a forward or backward stepwise manner. They have been used to assess the linearity of a variable being considered and the influence of individual cases.

The first, **Residual** is a component plus residual plot. The second, **Regression**, is a plot of added variables.

Residual

The Partial Residual plot, or component plus residual plot, is used to indicate whether you should make a transformation of an additional variable X that is a candidate variable to be included in a regression equation. It is also used to assess which transformation is useful and if influential cases exist.

The additional variable X_a is plotted on the horizontal axis. The residual, $e = Y - \hat{Y}$, (obtained when the additional X and the previously entered X s are used) is added to the slope coefficient associated with X_a times X_a , and then that sum is plotted on the vertical axis (i.e., the values of $(e + b_a X_a)$). The expected slope of the points is the regression coefficient of the entering or additional variable X_a . This plot can give an incorrect impression of the strength of the relationship between Y and the entering X if the multiple correlation is large (see Cook and Weisberg, 1982.)

Regression

The partial regression, or added variable plot, is useful for detecting lack of linearity in an additional X variable that is being considered as a candidate variable to be included in a regression model already containing one or more existing X variables. It is also used to assess influential cases.

When only the existing X s are included in the regression model, the residuals, $e = Y - \hat{Y}$ are plotted on the vertical axis. The residuals, $X_a - \hat{X}_a$, where a linear regression model is fitted with X_a as the dependent or outcome variable, and where the existing X s already in the regression model are the independent or predictor variables, are plotted on the horizontal axis. The intercept will be zero, and the slope of the points will be the slope of the regression coefficient of the additional X on Y . Ideally, you want the points to lie close to a straight line with no obvious outliers.

Custom Plots

This option allows you to plot any variables in the model, as well as any of the outlier diagnostic values in a scatterplot.

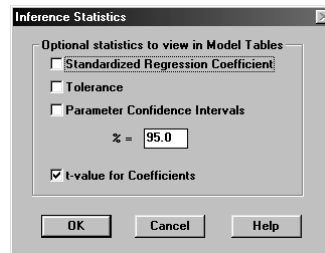
Diagnostics

Regression Diagnostics are listed by case for the variables in the model. The output first gives the case number and the value of the dependent variable so that you can relate the output back to the original datasheet. Next the output gives the predicted value of Y and the raw residual. Finally, the output provides statistics that are useful in finding outliers or influential cases. In the system, you can treat the datasheet containing these diagnostic data as if it were a standard datasheet. You can also obtain plots and descriptive statistics by using options in the menu. Here we will first define the various statistics listed and then review the plots and analyses that you can perform on them. For more detailed discussion of these topics see Chatterjee and Hadi (1989) or Cook and Weisberg (1982).

Options Menu

Inference Statistics

When you select **Inference Statistics** from the Multiple Regression Output window **Options** menu, a window is displayed that provides options to help you make inferences about the population regression model.



Standardized Regression Coefficients

Standardized regression coefficients (labeled Std_coeff) are the slopes of the regression model if X s and Y are standardized. Subtracting their respective means from each set of observations, and dividing by the respective standard deviation, standardizes the X s and Y . The standardized X s and Y have a mean of zero and a standard deviation of 1. The intercept term becomes zero.

The magnitude of the different standardized coefficients of the X s can be compared to determine the relative contribution of each X to the regression model. This is not the case with the unstandardized slope coefficients since the different magnitude and variability of the X s affects the size of the slope coefficients.

Standardized coefficients can be computed from unstandardized coefficients using the formula:

$$\text{Standardized } b_i = b_i (s_{X_i} / s_Y)$$

where s_{X_i} and s_Y are sample standard deviations of X_i and Y , respectively. For prediction, standardized slope coefficients must be used with the standardized data.

Tolerance - See the discussion earlier in this chapter.

Parameter Confidence Intervals

The 95% confidence limits for the slope coefficients are given by default. A 95% confidence limit has a 95% chance of covering the true population slope coefficient. In interpreting this interval, normality of the error terms is assumed. The $100(1-\alpha)\%$ confidence interval for β_i is:

$$b_i \pm t_{(1-\alpha/2, N-p-1)} s.e.(b_i)$$

where $t_{(1-\alpha/2, N-p-1)}$ is the upper $100(1-\alpha/2)\%$ quantile of a Student's t distribution, and $s.e.(b_i)$ is the standard error of b_i and there are $(N-p-1)$ degrees of freedom for a nonzero intercept model.

***t*-value for Coefficients**

The *t*-value tests the hypothesis that a population slope coefficient is zero. Normality is assumed. The *t*-test statistic is:

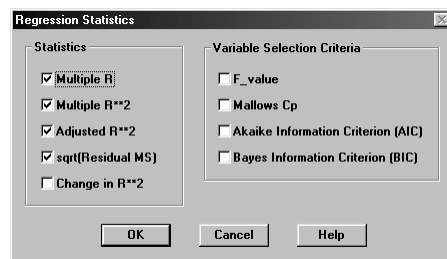
$$t = \frac{b_i}{s.e.(b_i)}$$

and is distributed as a Student's distribution with $(N-p-1)$ degrees of freedom for a nonzero intercept model.

When you perform this test for more than one β , the system does not give the joint significance level for all the slope coefficients. In such a case, you may want to use a Bonferroni adjustment. If you want a joint significance list for all of the regression parameters, use the *F* test under ANOVA.

Regression Statistics

Choosing the **Regression Statistics** option from the Multiple Regression **Options** menu allows you to obtain additional statistics and measures of model fit. You can choose options either for this analysis only, or to save as default. The window shown below is displayed.



Statistics

For the definitions of Multiple Correlation, Multiple Correlation Squared, Adjusted Correlation Squared, and Square Root of Residual Mean Square, see the section “Formulae” later in this chapter.

Change in Correlation Squared

Entering additional variables into a regression model will increase correlation squared (labeled Ch_R_sq). Conversely, removing variables will reduce correlation squared. The change in correlation squared will be larger for *X* variables that are independent or nearly independent of the other *X* variables already in the model, and for *X* variables that have higher explanatory power in predicting *Y*. Small changes in correlation squared indicate that there is little to be gained in adding a particular *X* variable if obtaining a better fitting sample regression model is your criterion.

Variable Selection Criterion

Sometimes theoretical considerations determine the variables which you should include in a regression model. In other instances, you may have no preconceived opinion on the choice of variables for inclusion in the model. Variable selection methods have been shown to be useful in

these latter exploratory situations. Variable selection procedures require a criterion for selecting variables and a stopping rule for deciding when to quit entering variables.

F-value

The F -value is used to assess the significance of the regression parameters. It is the ratio of the regression mean square to the residual mean square. See the ANOVA table discussion earlier in this chapter for more information.

Mallows Cp

The Mallows Cp criterion is a widely used criterion. Variables are selected that result in a small C_p . The criterion can be expressed as:

$$C_p = (N - p - 1) \left(\frac{\text{RMS}_{(p)}}{s_e^2} - 1 \right) + (p + 1)$$

where $\text{RMS}_{(p)}$ is the residual mean square based on p selected predictor variables, and s_e^2 is the residual mean square including all predictor variables in the pool.

NOTE: The second term $(p+1)$ will increase in size as more variables are added, but the first part of the formula will decrease and finally become zero when all the variables are entered. For a derivation of C_p , see Daniel and Wood (1971.)

A plot of C_p on the vertical axis versus p on the horizontal axis will tend to decrease as more variables are added, and then it starts to increase.

One suggested rule is to choose the set of variables that minimizes C_p (see Mallows, 1973.)

Akaike Information Criterion (AIC)

The Akaike information criterion is a recommended selection criterion. It has been shown to be asymptotically equivalent under the null hypothesis to C_p and the risk based on C_p will converge to the risk based on AIC (see Nishi, 1984.) It is given as:

$$\text{AIC} = \text{RSS}_{(p)} / s_e^2 + 2(p + 1)$$

where:

p is the number of X variables. Assuming a nonzero intercept, $\text{RSS}_{(p)}$ is the residual sum of squares with p variables in the model, and s_e^2 is the residual mean square based on a full regression model with all X variables included.

Bayes Information Criterion (BIC)

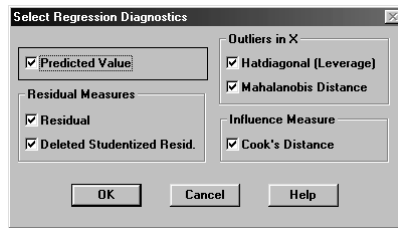
Bayes information criterion (BIC) is another selection criterion. Its formula is given by

$$\text{BIC} = \text{RSS}_{(p)} / s_e^2 + (p + 1)\log(N)$$

where p is the number of X variables. Assuming a nonzero intercept, $\text{RSS}_{(p)}$ is the residual sum of squares with p variables in the model, and s_e^2 is the residual mean square based on a full regression model with all X variables included.

Diagnostics

Choosing **Diagnostics** from the Multiple Regression **Options** menu allows you to select from a list of diagnostics for detection of outliers. You can decide which diagnostics you want to see listed by case in a datasheet format. Diagnostics included predicted values, residuals, deleted Studentized residuals, hat diagonal elements, Mahalanobis' distances, and Cook's distances.

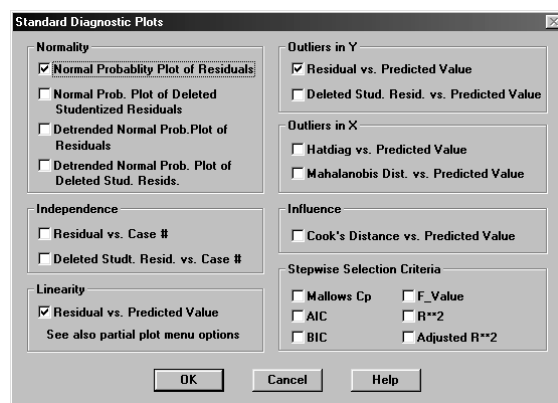


For definitions of:

Predicted Value, Residual, Studentized Residual (deleted), Hat Diagonal, Mahalanobis' Distance, and Cook's Distance, see the section "Formulae" later in this chapter.

Diagnostic Plots

Choosing the **Diagnostic Plots** option from the Multiple Regression **Options** menu displays a window showing commonly used diagnostic plots by function.



NOTE: You can obtain additional plots by selecting the View Menu and then selecting Plots. In addition to the plots given here, custom plots and partial plots are available. You can also append your diagnostic variables back to the datasheet and more plot options will be available. When you do so, the appended variables are unlinked from the datasheet.

Normality

When tests of hypotheses or confidence limits are computed, the error terms, e , are assumed to be independently, normally distributed with mean zero and constant variance, σ_e^2 . Residuals from the sample regression model are used to compute s_e^2 , which is a point estimate of σ_e^2 .

Thus, if you want to test hypotheses or construct confidence limits, you will use residuals, the estimates of the error terms, to assess normality. There are some problems with this approach. The sample residuals are not independent, and they do not have the same variance unless special conditions are met (see Chatterjee and Hadi, 1989). However, the sample residuals are still useful in practice to discover lack of normality.

Normal Probability Plot of Residuals/Deleted Studentized Residuals

The normal probability plot is used to show graphically whether the distribution of sample residuals or deleted studentized residuals is normally distributed with a constant variance or not. The numerical values of the residuals are plotted on the horizontal axis and the expected normal values based on ranks are plotted on the vertical axis. If residuals are normally distributed, the points should lie on a straight line.

Deleted studentized residuals are recommended for use in plotting normal probability plots over sample residuals. That is because deleted studentized residuals have a t distribution if the usual normality assumption holds (see Cook and Weisberg, 1982), although the difference may not be very noticeable.

Detrended Normal Probability Plot of Residuals/Deleted Studentized Residuals

Detrended normal probability plots are similar to normal probability plots. However, in the detrended normal probability plot, the slope of the plot is made to be zero and the vertical scale is expanded so that you can easily see any deviations. If data are normally distributed the points will lie in a horizontal straight line clustered about zero on the vertical axis.

Deleted studentized residuals are recommended for use in plotting detrended normal probability plots over sample residuals. This is because deleted studentized residuals have a t distribution if the usual normality assumptions hold (see Cook and Weisberg, 1982), although the difference may not be very noticeable.

Independence

Error terms are assumed to be independent, but their estimates, the residuals, are not. Nevertheless, a plot of residual versus case number has been shown to be useful in practice in detecting appreciable departures from independence. These plots are useful in assessing whether trends in the magnitude or sign of the residuals are occurring over cases, or if clumping of case residuals has occurred at some point.

Residual versus Case #/Studentized Residuals (deleted) versus Case #

A plot of residual versus case number provides a quick and simple method of checking for gross trends in the residuals from successive cases. Often, successive cases can serve as a proxy for time. Thus these plots serve to show whether the regression model fits the early cases in a fashion similar to the later cases. Since residuals do not have precisely the same variance and are not, strictly speaking, independent, this plot should be examined for general trends.

A plot of the deleted studentized residual versus the case number provides the a quick and simple method of checking for gross trends in the residuals for successive cases.

Often, successive cases can serve as a proxy for time. These plots serve to show whether the regression model fits the early cases in a fashion similar to the later cases.

Linearity

Usually, in fitting a regression model you will first fit an equation where the Y variable and X variables are used in the form in which they were measured. When this is done, you are using a

model where the X variables are linearly related to the Y variable. Sometimes this model is not adequate.

A plot of the residual versus the predicted Y or \hat{Y} provides a quick way of assessing the fit of a model. If problems exist, one common procedure is to consider making transformations on some or all of the X s or on Y . The resulting model is still linear in the parameters (slope coefficients), as required, if a linear regression model is to be fitted. If no suitable model can be found that is linear in the parameters, then non-linear regression may be necessary.

You can also use a Box-Cox transformation on the Y variable. In general, many analysts prefer to make transformations on the X s rather than on Y . Two reasons for this are: 1.) only some of the X s may not be linearly related to Y ; and 2.) if you want to make predictions, the results are easier to interpret if Y is left in its original units. Nevertheless, there are times when transforming the Y variable is a valid procedure.

If you are considering the transformation of some or all of the X s, then you should make additional plots. A common first step is to determine whether the various X s are approximately normally distributed by plotting them univariately using histograms, boxplots, or normal probability plots. The system's simple linear regression option is recommended. Simple linear regression plots each X variable against Y , one at a time, then partial plots should be considered (go to the View Menu and click on Plots). Also, it is recommended that residuals from the multiple regression model be plotted against X (see Custom Plots). This univariate approach will work well if the predictor variables are uncorrelated with each other. Caution must be taken when the X s are correlated.

Residual versus Predicted Value

This is one of the most widely used diagnostic plots. Residuals are plotted on the vertical axis, and predicted values of Y or \hat{Y} from the regression equation are plotted on the horizontal axis. If a linear relationship holds, and the residuals at least approximately follow a normal distribution, then the plot of the points should cluster around a straight line with a zero slope and approximately equal numbers of points above and below the residual value of zero.

If the points appear to follow a curve, then transformations of either the Y or some or all of the X s may be useful.

This same plot is used in determining whether outliers exist and as a rough check for equality of variances of the error terms.

Outliers

Five scatterplots are provided to help you find possible outliers. The plots are: residual versus predicted values of Y ; deleted Studentized residuals versus predicted values of Y ; hat diagonal elements versus predicted values of Y ; Mahalanobis' distance versus predicted values of Y ; and Cook's distance versus predicted values of Y .

Residuals versus Predicted Values

The plot of the residuals against predicted values is often used to find outliers in Y . These residuals do not have equal variance, so their interpretation is not as straightforward as the interpretation of deleted Studentized residuals. However, in many instances the pattern of the points will be similar.

Some users find residuals easier to interpret, since they are in the original units. Cases at extreme points from a horizontal line drawn through zero on the vertical axis are possible outliers.

If they are also at the extremes of \hat{Y} , then these cases are likely to have a greater effect on the slope coefficients.

Studentized Residuals (deleted) versus Predicted Value

Deleted Studentized residuals in Y follow a Student t distribution with $(N-p-2)$ degrees of freedom. Cases which are more than two and a half units from a horizontal line drawn through zero in the vertical axis are potential outliers in Y .

Hat Diagonal versus Predicted Value

Diagonal elements of that matrix h_{ii} or leverage statistics measures outliers in the X s. Plotted values of h_{ii} that are greater than twice the mean are sometimes examined as possible outliers in X . See Hat Diagonal.

Mahalanobis' Distance versus Predicted Value

Mahalanobis' distance is another measure of outliers in the X s. Mahalanobis' distance is a measure of the distance from each case (using only the X variables in the model) to the mean of all the X s in the model. Large distance values indicate outliers in the X s. Plotted values of Mahalanobis' distance versus predicted values of Y are used to find cases that have a high leverage. See Mahalanobis' distance.

Cook's Distance versus Predicted Value

Cook's distance is a measure of influence on the regression coefficients that combines the effects of outliers in Y and in X s. Plots of Cook's distance against predicted value of Y is used to find cases that can have a sizable effect on the regression coefficients. Cook's distance is always positive. See Cook's Distance.

Variable Selection Criteria

Different selection criteria such as Multiple Correlation Squared, and Adjusted Correlation Squared can be plotted against p (the number of predictor variables in the model). Step number labels the points so that you can trace back a particular model from the Stepwise Regression Summaries window.

Rerun Menu

The **Rerun** menu in the Output window provides two options where you can select a new Y variable, either to rerun the current regression analysis, or specify a new analysis.

Formulae

Predicted Value

The predicted value of Y or \hat{Y} (labeled as Predicted) is obtained from a regression equation for each case by entering the values of the X variables into the fitted regression equation.

Residual

Residual (labeled as Residual) is the difference between Y and its predicted value, \hat{Y} . Sometimes these are called raw residuals to distinguish them from residuals that have been transformed. The average value of the residuals is zero. Residual values that are large in absolute value are an indication of cases that do not lie close to the hyperplane. These are used both to determine whether the model fits and to detect outliers.

Studentized Residual (deleted)

This is often called the deleted Studentized Residual or externally Studentized Residual (labeled as DelStRes). The residual is computed to the i th case from a regression equation with the i th case deleted. Hence, the case for which a residual is found is not included in the estimation of regression parameters. If a case is an extreme outlier, then removing it before computing the residual removes the effect of that outlier from the estimate of the standard error. The Deleted Studentized residual is defined as:

$$\text{Deleted Studentized Residual} = \frac{e_i}{s_{(i)}(1 - h_{ii})^{1/2}}$$

where e_i is the i th raw residual from a regression equation, $s_{(i)}$ is an estimate of the standard error about the regression hyperplane with the i th case omitted, and h_{ii} is the element in the diagonal of the hat matrix, $X(X'X)^{-1}X'$ given in the outliers in the next column.

The value of the residual is divided by its standard error, so values that are larger than 3 in absolute value are candidates for consideration as outliers. If the vector of error terms is normally distributed, ($\epsilon \sim N(0, \sigma_\epsilon^2 I)$), then the deleted Studentized Residuals are distributed as a t distribution with $(N-p-2)$ degrees of freedom, assuming you used an intercept model. For further discussion on interpreting these and other residuals, see Chatterjee and Hadi (1989) or Cook and Weisberg (1982.)

Hat Diagonal

Diagonal elements of the hat matrix, $X(X'X)^{-1}X'$, h_{ii} (labeled as HatDiag), are statistics that are widely used to find outliers in the X s. In general, measures that are used to find outliers in X are called leverage statistics. Large values of h_{ii} are an indication of possible outliers in the X s. The size of h_{ii} is limited to the range of $1/N$ to 1. The average value of h_{ii} is p/N .

The h_{ii} for the i th observation tells how much an observation contributes to the estimation of regression parameters. Observations with large leverages have the potential for having a large effect on the regression coefficients. If they also have a large deleted Studentized residual, then they should be considered as possible outliers.

For p variables, Chatterjee and Hadi (1989) recommend investigating values of $h_{ii} > 2.5p/N$ if $p > 2$ and $h_{ii} > 2p/N$ if $p > 6$. The numerical value of h_{ii} is computed from:

$$h_{ii} = x_i(X'X)^{-1}x_i$$

where x_i is the i th row of X , including one row for the intercept.

Mahalanobis' Distance

The leverage of a case can also be measured by its Mahalanobis' distance (labeled as MahDist). This statistic measures how distant a case is from the mean of the set of X s and is used to find outliers in X . It is computed using only the X variables in the regression model. Observations with large leverage have the potential for having a large effect on regression coefficients.

The Mahalanobis' distance is related to h_{ii} by the following formula:

$$\text{Mahalanobis' Distance} = (N-1)(h_{ii} - 1/N).$$

Cook's distance

Determining the influence of each case on the regression coefficients is a direct method of detecting outliers in a regression model. Cook derived a function called Cook's distance (labeled as CookDist) which gives a scaled distance between the value of the regression estimates when all the cases are present, and the value when the cases are omitted one at a time, $b_{(-i)}$. Cases with a large Cook's distance are possible influential observations.

Cook's distance for the i th case is defined as:

$$C_i = \frac{(b - b_{(-i)})'(X'X)(b - b_{(-i)})}{(p+1)S_e^2}$$

Where p is the number of X variables, and S_e^2 is an estimate of the σ_e^2 ($S_e^2 = \text{ResMS}$). Values of C_i will increase with increasing difference between regression coefficients for all N and when the i th case is removed. Subsequently, it has been found that Cook's distance can be written as:

$$C_i = \frac{e_{(i)}^2 h_{ii}}{p S_e^2}$$

where $e_{(i)}$ is the i th residual computed from Y minus the predicted Y computed from the regression model with the i th case deleted. It can be seen from this formula that Cook's distance takes into account both leverage h_{ii} and the size of the residual Y .

Strictly speaking, C_i does not have an F distribution, but an F distribution has been suggested as a way of deciding which cases should be considered as possible outliers. Cases for which the value of Cook's distance is greater than the 95th percentile of the F distribution with $(p+1)$ and $(N-p-1)$ degrees of freedom may be considered for further investigation.

N = Effective number of cases used in fitting the regression model.

Multiple Correlation

Sample multiple correlation is the Pearson r (product moment correlation) between Y and its predicted value, \hat{Y} , obtained by fitting a regression model. The sample multiple correlation is always nonnegative and lies between 0.0 and 1.0. It is a measure of the linear relationship between Y and the set of X variables. When the sample multiple correlation is close to 1.0, then

the relationship is highly linear. When it is close to 0.0, the hyperplane barely fits the Y variables better than the mean of Y .

Population multiple correlation is defined as the square root of:

$$\rho_{Y:X_1, \dots, X_p}^2 = 1 - (\sigma_\epsilon^2 / \sigma_Y^2)$$

Where σ_ϵ^2 is the variance of Y about the population regression hyperplane, and σ_Y^2 is the variance of a population of all possible Y s. When σ_ϵ^2 is small relative to σ_Y^2 , then the multiple correlation will be large (close to 1.0.) The sample multiple correlation (labeled as `Multiple_R`) is a biased estimate of the population multiple correlation and tends to overestimate its value.

When you request the zero-intercept model, the sample multiple correlation is based upon uncorrected sums of squares (*i.e.*, the means are not removed).

Multiple Correlation Squared

Multiple correlation squared, R^2 , (labeled as `R_sq`) is often called the coefficient of determination. Its magnitude is an indication of the reduction in variance of Y that has been achieved by fitting the regression model. See Multiple Correlation.

Adjusted Correlation Squared

Adjusted correlation squared (labeled as `Adj_R_sq`) has less bias in estimating the population multiple correlation squared than does the multiple correlation squared. Its value can be negative. For large N and a small number of X variables in the regression model, it will be close to multiple R^2 in value. The multiple correlation squared is adjusted by taking into account the number of variables (p) in the model.

$$\text{Adjusted correlation squared} = R^2 - p(1 - R^2)/(N - p - 1).$$

As more X variables are added to the model, the value of R^2 will either increase or stay the same, but the value of adjusted R^2 may decrease. Some analysts will stop adding variables to the model when the adjusted R^2 is at its maximum.

Partial Correlation

Partial correlation (labeled `Partial_r`) is used to measure the correlation between two variables with the linear effect of one or more other variables removed. Suppose you have two X variables whose partial correlations you wish to obtain by partialling out the linear effects of three other X variables. You could first fit a regression plane predicting one of the two X variables from the three X variables. Second, you could fit a second plane predicting the second X variable from the same three X variables. You could finally get the residuals from each of these planes and find the correlation of the two sets of residuals. This correlation is the desired partial correlation. Alternative formulas are used in the system, but the results are essentially the same.

Partial correlation is useful in forward and stepwise regressions in determining the effects of adding another variable to the regression model, given that you already have some variables entered. The linear effect of variables already in the equation is partialled out.

Square Root of Residual Mean Square

The square root of the residual mean square (labeled `sqrt (ResMs)`) is the square root of the variance of Y about the regression hyperplane.

$$\text{Sqrt(ResMS)} = \sqrt{\sum_{j=1}^n (y_j - \hat{y}_j)^2 / (N - p - 1)}.$$

It is an estimate of the error standard deviation and is used in confidence interval construction.

Regression Coefficients

Regression Coefficients (labeled as `Coeff`) are computed for each variable in the model. Unless the X variables are uncorrelated, the numerical value of the regression coefficient for a given variable will change as different variables are entered into the model. They are computed from:

$$b = (X'X)^{-1}X'Y$$

Standard Error of the Coefficients

Standard error of the coefficients (labeled as `SE_coeff`) is the standard deviation of the coefficient estimates. The normality assumption is given by:

$$\left(\varepsilon_j \sim N(0, \sigma_\varepsilon^2) \right)$$

If the normality assumption holds, the standard error of the coefficients can be used to perform a t - or an F -test where the null hypothesis is zero slope. The squared standard error (or the variance) of the coefficient estimates is computed by multiplying the residual mean square by the appropriate diagonal element of the matrix $(X'X)^{-1}$.

F-to-Enter

Specifies the minimum F -to-Enter value that must be exceeded by the computed F -to-Enter value for a candidate variable to enter in the automatic forward stepwise method. Suppose that p predictor variables are already entered in the model. If you enter the candidate variable, there will be $p+1$ variables entered. The computed F -to-Enter is:

$$F = [RSS_{(p)} - RSS_{(p+1)}] / [RSS_{(p+1)} / (N - p - 2)]$$

where $RSS_{(p)}$ denotes the residual sum of squares with p variables in the model. The degrees of freedom are 1 (the difference in the number of parameters between the two models) and $(N-p-2)$ since there are $(p+1)$ slope coefficients plus the intercept. The usual tabled F -values from an F distribution do not apply for statistical testing in stepwise selection, since the system is selecting the best variable at each step. The appropriate critical value is a function of the number of cases, the number of variables, and the correlation structure of the X variables.

A commonly used cutoff point for F -to-Enter values is an F -value equivalent to a p -value of 0.10 to 0.25 when the purpose of fitting the regression model is predictive. Useful values to use in practice are a minimum of F -to-enter value 2.07 (see Bendel and Afifi, 1977) or for a smaller p -value, a default F -value of 4.0.

For the zero intercept model, $RSS_{(p)}$ denotes the uncorrected residual sums of squares, and the denominator degrees of freedom is $(N-p-1)$.

F-to-Remove

The maximum F -to-remove is an F -value to which the computed F -value must be compared in order to see if a variable should be removed from the model. The variable with the smallest computed F -value is removed first if it is less than the maximum F -to-remove. Variables are removed until there are no variables in the model with a computed F less than the maximum F -to-remove value.

The computed F -to-remove value for a candidate variable X_i is :

$$F = [RSS_{(p-1)} - RSS_{(p)}] / [RSS_{(p)} / (N - p - 1)]$$

where $RSS_{(p)}$ denotes the residual sum of squares with p variables in the model. The degrees of freedom are 1 and $(N-p-1)$ for a nonuser intercept model.

F -to-remove tests the relative importance of variables already in the equation and removes variables in the order of their contribution to the model.

In forward stepwise regression, if the F -to-remove value is chosen to be much smaller numerically than the F -to-enter value, then variables that are already entered will tend not to be removed. The procedure reverts to what is called forward selection rather than forward stepwise regression.

NOTE: The computed F -to-remove value of a variable following its inclusion in forward stepwise at step $(k + 1)$ is precisely its computed F -to-enter value at step k .

The usual tabled F -value should not be interpreted as giving the usual p -values since numerous tests are made. See the discussion on F -to-Enter.

4. Tables - Frequency Analysis

USING FREQUENCY ANALYSIS

FREQUENCY ANALYSIS OUTPUT OPTIONS

TABLES

TESTS

Introduction

The system provides two-way tables of frequencies, row percents, column percents, overall percents, expected values under independence, and differences between the observed and expected frequencies. Components of chi-square and standard deviates are also provided for each cell in a table. Four measures of association are available (refer to Tables in this Chapter), and five tests are available (refer to Tests in this Chapter).

	CC1	CC2	CC3	Total
RC1	n_{11}	n_{12}	n_{13}	$n_{1.}$
RC2	n_{21}	n_{22}	n_{23}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

You can construct two-way tables from a datasheet or from the frequency table editor. Tabulated data can be read into a datasheet with a case corresponding to a unique cell in the table and a case frequency variable defined.

In general, n_{ij} refers to the cell frequency in the i th row ($i=1,\dots,r$) and the j th column ($j=1,\dots,c$), while $n_{i.}$ refers to the sum of the cell frequencies in the i th row, and $n_{.j}$ refers to the sum of the cell frequencies in the j th column, and N is the total sample size.

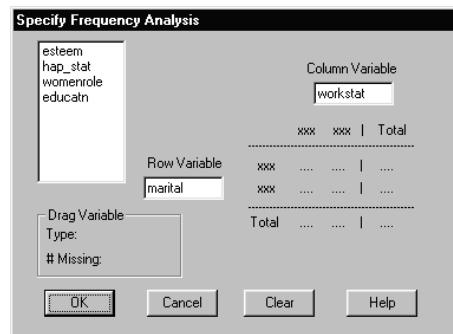
In some cases, cell, row total, and column proportions are used, and are denoted as $p_{ij} (=n_{ij}/N)$, $p_{i.} (=n_{i.}/N)$, and $p_{.j} (=n_{.j}/N)$, respectively.

Under certain assumptions, such as independence, and given an estimation procedure, estimates of cell frequencies will be denoted by e_{ij} .

Using Frequency Analysis

To perform Frequency Analysis, execute the following steps:

1. Choose **Tables** from the datasheet **Analyze** menu to display the window shown below:



2. Drag and drop the Row and Column variables to their respective datafields (the selected variables must have more than one group). Press the **OK** button.
3. The Output window showing the default output options is displayed.

Frequency Output: FIDELL - marital and workstat

File Edit View Options Format Window Help

Arial 10 B I U

TESTS

	Value	df	p-value
Pearson's Chi-square	33.0835	2	6.546E-08
Likelihood Ratio Chi-square	36.7352	2	3.879E-09

TABLES

Table of Counts:

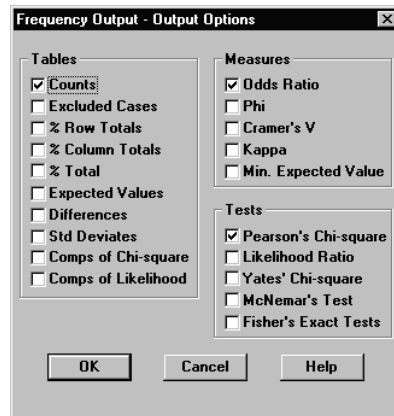
	paidwork	hap_home	unhap_hom	Total
single	77	8	18	103
married	169	129	64	362
Total	246	137	82	465

NUM Col: 0 Line: 64 Page: 1

Frequency Analysis Output Options

The output options - Frequency Tables, Measures, and Tests are described in the following sections.

1. From the Frequency Tables Output window **Options** menu, choose **Output** to display the window shown below.
2. You can check additional output options for Tables, Measures and Tests, then press the **OK** button to view the results in a new Output window.



Tables

Tables of percents are given so that row totals are 100% (% Row Total), column totals are 100% (% Column Total), and overall N is 100% (% Total). Other statistics are included to assist in interpreting the effects of individual cells. See expected values, differences, standard deviates, and components of Chi-square and likelihood ratio G^2 .

Four measures of association are given:

- ◆ Odds ratio
- ◆ Phi coefficient
- ◆ Cramer's V
- ◆ Kappa statistic

Odds ratio is one of the more commonly used measures, but construction of a unique set is limited to two by two tables (see Agresti, 1990). The Odds ratio is given only for two by two tables. Phi coefficient and Cramer's V are functions of chi-square, and their significance can be determined using the *p*-value from the chi-square test. Kappa statistic is used as a measure of interrater agreement (see Fleiss, 1981). It requires a table with an equal number of rows and columns.

In addition to measures of association, the system provides descriptive information to assist in data interpretation. The minimum expected value of chi-square over all cells is printed out, and you can determine whether this value is too small. In such cases, results of the chi-square test should be interpreted with caution.

Tests

Five tests are available:

- ◆ The Pearson chi-square test
- ◆ The chi-square test with Yates' correction
- ◆ Likelihood ratio G^2 test
- ◆ McNemar's test of symmetry (the number of rows and columns must be equal)
- ◆ Fisher's exact test for 2x2 tables

Expected Values

The expected value for each cell (e_{ij}) is computed assuming independence and the minimum value is displayed. The expected value is defined as:

$$e_{ij} = n_{i.}n_{.j}/N.$$

The minimum value is included to provide you with information for evaluating whether or not it is safe to assume that the reported p -value based on a chi-square distribution can be applied to your data set.

In general, if the minimum expected value is greater than 2, or no more than 20% of the expected values are less than 5 with a minimum of 1, the chi-square approximation of the sampling distribution of the test statistic is adequate (see Dixon and Massey, 1983). If you have a very small minimum expected value or numerous small expected values, you might wish to combine some rows or columns or use Fisher's exact test.

Differences

The difference between Expected Values under independence and observed frequency is given for each cell:

$$\text{Difference} = n_{ij} - e_{ij}$$

Large absolute differences indicate violation of the assumption of independence.

Components of Chi-square

Components of chi-square, $(n_{ij} - e_{ij})^2 / e_{ij}$, are computed for each cell. Large component values are indicative of departure from the independence assumption.

Standardized Deviates

Standardized deviates are the square root of the components of Chi square.

Components of Likelihood Ratio G-square

Components of likelihood ratio G^2 are computed as: $2n_{ij}\ln(n_{ij}/e_{ij})$.

Odds Ratio

Odds ratio (or cross-product ratio or OR) may be used to measure association in two by two (two rows and two columns) tables. Odds ratio is calculated by:

$$OR = n_{11}n_{22} / (n_{12}n_{21}).$$

If we think of the column totals as fixed, then n_{11}/n_{21} is the odds of being in the first row conditional on being in the first column. Then n_{12}/n_{22} is the same odds for the second column. OR is then the ratio of these odds.

OR can take on values from 0 to infinity, and is equal to 1 if there is no association. Values less than 1 indicate a negative association and those greater than 1 a positive association. OR is not symmetric about 1.

OR is invariant under interchange of rows and columns. It is also invariant under row and column multiplication.

In addition to odds ratio, the natural logarithm of OR [Ln(OR)] is given. Ln(OR) is symmetric about zero and runs from minus infinity to plus infinity. The system provides Ln(OR) along with its asymptotic standard error, labeled S.E./Ha. Asymptotic standard error is computed as follows:

$$S.E./Ha = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}.$$

This can be used to compute confidence limits for Ln(OR) and the confidence interval for OR may be derived by exponentiating the confidence limits of Ln(OR).

Under the null hypothesis of no association, the asymptotic standard error (A.S.E.) is calculated differently:

$$A.S.E. = \sqrt{N^3 / [(n_{11} + n_{12})(n_{11} + n_{21})(n_{12} + n_{22})(n_{21} + n_{22})]}.$$

The normal-based test statistic value (labeled as $z_value = Ln(OR)/A.S.E.$) can be used to test for the null hypothesis of no association. You will reject the null hypothesis if the associated p -value is less than a specified level of significance.

Phi coefficient

Phi coefficient is a measure of association defined as the square root of the chi-square value χ^2 divided by N .

$$Phi = \sqrt{\chi^2 / N}.$$

For a two by two table, it varies between 0 and 1 and its square is equivalent to the square of the product-moment correlation of two binary variables. A sign has been included in the system phi coefficient (for two by two tables) so that you can see whether the association is positive or negative:

$$Phi = (n_{11}n_{22} - n_{12}n_{21}) / \sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

In general, the maximum value of phi squared is equal to the minimum of the number of rows minus 1 and the number of columns minus 1. Hence, for a table with more than two rows or columns, it cannot equal 1. The system provides the maximum possible value of phi coefficient for your table structure to help you evaluate the value of phi coefficient. See Fleiss (1981) for a summary of the pros and cons of using this statistic.

Cramer's V

Cramer's V is a measure of association defined as:

The square root of the chi-square value, χ^2 , divided by N times the minimum of the number of rows minus 1 and the number of columns minus 1.

$$\text{Cramer's } V = \sqrt{\chi^2 / [N \min(r-1, c-1)]}.$$

For a two by two table Cramer's V is equal to the absolute value of the phi coefficient. For a larger table, Cramer's V can be equal to 1. It is sometimes used when investigators wish to compare tables with different numbers of rows or columns and different total sample sizes.

Kappa statistic

Kappa statistic is a measure of inter-rater agreement or reliability that removes the effect of chance agreement (see Fleiss, 1981 for examples). It can only be computed on tables that have the same number of rows as columns, say k.

For example, suppose two raters assign each case into one of the k categories, the results for one rater are given along the rows, and the results for the other rater along the columns. Frequency numbers in the diagonal cells are cases of agreements between the two raters.

The formula for Kappa statistic is:

$$\text{Kappa} = (p_o - p_e) / (1 - p_e),$$

where:

$$p_o = \sum_{i=1}^k n_{ii} / N \text{ (observed proportion of agreement), and}$$

$$p_e = \sum_{i=1}^k n_{i.} n_{.i} / N^2 \text{ (expected proportion of agreement).}$$

The estimated standard error of Kappa statistic (labeled as SE/Ha) is computed as

$$\text{SE/Ha} = \sqrt{(A + B - C) / N} / (1 - p_e),$$

where:

$$A = \sum p_{ii} [1 - (p_{i.} + p_{.i})(1 - \text{Kappa})]^2,$$

$$B = (1 - \text{Kappa})^2 \sum \sum_{i \neq j} p_{ij} (p_{i.} + p_{.j})^2,$$

$$C = [\text{Kappa} - p_e (1 - \text{Kappa})]^2.$$

The preceding standard error is used in constructing a confidence interval for the population value of Kappa. The standard error of Kappa is also used in hypothesis tests where Kappa is hypothesised to be some nonzero value.

Under the null hypothesis that the underlying value of Kappa is zero or no agreement, the standard error used in calculating the normal - based test statistic (labeled z_value) is:

$$A.S.E. = \frac{1}{(1 - p_e)\sqrt{N}} \sqrt{p_e + p_e^2 - \sum p_{i.} p_{.i} (p_{i.} + p_{.i})}$$

The standard error may not be accurate for small sample sizes (see Wickens, 1989).

Chi-square test

You can use the Pearson chi-square test for either a test of independence between the row and column variables or a test of homogeneity. In the test for independence, the total sample size is fixed. In the test for homogeneity, the row (or column) totals are fixed. For example, if you had three treatment groups and had assigned a fixed number of patients to each treatment, then you could perform a test of homogeneity. In that case you are testing the hypothesis that the outcomes of the three treatments are equal.

In the test for independence, the sampling distribution of cell frequencies is a multinomial, and for homogeneity it is independent (sometimes-called product) multinomials (see Bishop *et al.*, 1975 for discussion of the sampling distributions.)

The familiar chi-square statistic, χ^2 , is computed as:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c (n_{ij} - e_{ij})^2 / e_{ij},$$

where: $e_{ij} = n_{i.} n_{.j} / N$.

A large value of chi-square and/or a small p -value signifies that there is a small chance of making an error if you reject the null hypothesis.

If a chi-square distribution is to be a good approximation to the multinomial distribution, there must be few expected values that are too small. See minimum expected values.

The degrees of freedom are the number of rows minus 1 times the number of columns minus 1:

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

Likelihood Ratio Chi-square

Likelihood ratio chi-square test is a general maximum likelihood-based hypothesis test. It is useful for comparing nested models. In this test, the likelihood under the null hypothesis and the likelihood under a general model are maximized. The ratio of the maximum likelihoods (which cannot exceed 1) is called λ . It has been shown that $-2\ln(\lambda)$ has a limiting chi-square distribution as N approaches infinity.

Under the null hypothesis of independence, the test statistic is:

$$G^2 = 2 \sum \sum n_{ij} \ln(n_{ij} / e_{ij})$$

and is distributed as a chi-square random variable with degrees of freedom equal to the number of rows minus 1 times the number of columns minus 1 ($df=(r-1)(c-1)$) (See Agresti 1990). The larger the value of likelihood ratio G^2 , the lower the likelihood that you will make an error in rejecting the null hypothesis.

When independence holds, χ^2 and likelihood ratio G^2 are asymptotically equivalent and, in most cases, their numerical values and resulting p -values will be quite similar. The question of which one to use arises.

It has been shown that chi-square is valid with smaller sample sizes and sparser tables than the likelihood ratio chi-square (again, see Agresti, 1990, section 7.73) but no single rule appears to

easily differentiate between the two statistics. On the other hand, likelihood ratio tests are widely used and have attractive properties (see section 4.2 in Bishop *et al.*, 1975.)

Yates' Corrected Chi-square

For two by two tables, one-half is subtracted from the absolute value of $(n_{ij} - e_{ij})$ before squaring it and computing chi-square (see chi-square for computation.) This “correction” was proposed by Yates as a method of correcting for continuity (Yates, 1934.)

The Yates chi-square will always be smaller than the Pearson chi-square. Its use has been debated in several articles. It is regarded as a conservative test in the sense that the level of significance of the test is lower than the stated level.

McNemar's Test of Symmetry

McNemar's test can be used for paired or dependent samples when the outcome variable is categorical data (nominal, ordinal or continuous data that has been grouped). The number of rows must equal the number of columns. This test is often used when the same cases are measured at two different time periods.

The null hypothesis being tested is of population symmetry. That is, the population proportions are the same in symmetrically located cells about the diagonal. The test statistic is:

$$\chi^2_{MC} = \sum \sum_{i < j} (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji})$$

and is distributed as a chi-square random variable with degrees of freedom equal to the number of rows times the number of rows minus 1, all divided by 2 (*i.e.*, $df = r(r-1)/2$).

Fisher's Exact Test

Fisher's exact test is implemented for two by two tables only. It is an exact test that does not require approximation. Hence, it is widely used for small sample sizes when the expected values in one or more of the cells are too small for either the chi-square, or the likelihood ratio G^2 test. In general, the system computes the probability of all possible ways in which the data could come out more deviant than it did. Under the null hypothesis of independence, and under Poisson or multinomial sampling, a hypergeometric distribution is obtained if we condition on the totals in both margins.

The test is performed by summing the hypergeometric probabilities for outcomes as deviant or more deviant than that obtained. One-tailed and two-tailed p -values are provided. In using the one-tailed p -value, you must verify whether the tail chosen corresponds to the tail appropriate for your test. It is recommended that a two-tailed test be used if you want the results to be comparable to the chi-square test. The exact p -values for a given sample are not continuous since only certain probabilities can result. In calculating for one-tailed probability, the two by two table is rearranged as follows:

a	b
c	d

and so that $ad \leq bc$ and $a \leq d$.

The one-tailed probability is the sum of observed values less than or equal to a :

$$p\text{-value(1-tail)} = \sum_{x=0}^a \frac{n_1!n_2!n_1!n_2!}{N!(a-x)!(b+x)!(c+x)!(d-x)!}$$

where:

$$y! = \prod_{i=1}^y i = 1 \times 2 \times \dots \times y \text{ (factorial) and } 0! = 1.$$

The other tail probability is obtained by adding probabilities of observing a cell value that is greater than a such that the individual probability does not exceed the probability of observing a . This tail probability is then added to the one-tailed probability to obtain the two-tailed p -value.

5. *t*- and Non-parametric Tests

USING *t* AND NON-PARAMETRIC TESTS

OUTPUT OPTIONS

Introduction

You can perform tests on location parameters for a single sample or for two samples. In the two sample analysis case, samples may either be independent of, or dependent (paired data) on each other. Two independent samples may arise either from an experiment where there is random assignment to two treatment groups, or from a survey where two groups are being compared. Paired or dependent samples are often two measurements on the same subject done at two time periods, or under different conditions. Dependent samples might also be two distinct subjects who are paired because they are similar on some other characteristics.

The *t*-tests output is displayed, the Non-parametric tests output is displayed next followed by results from a robust test (if requested).

Notation

The following definitions will be used in this chapter:

μ_i is the population mean of variable X_i , $i=1, 2$. If there is only one population of interest, the subscript will be omitted (e.g., μ).

n_i is the sample size from the i th population (group or stratum).

N is the total sample size.

\bar{x}_i is the sample mean of the i th group.

s_i^2 and s_i are the sample variance and sample standard deviation of the i th group and:

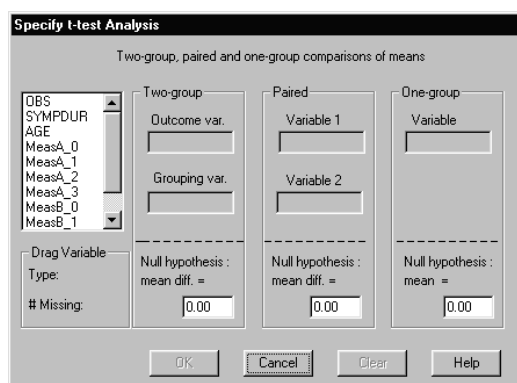
$t_{(1-\alpha/2, df)}$ is the $(1-\alpha/2)100$ percentile of a Student's *t* random variable with *df* degrees of freedom and $0 < \alpha < 1$. That is, the density area to the right of $t_{(1-\alpha/2, df)}$ is $\alpha/2$.

Using *t*- and Non-parameteric Tests

Select ***t* and Nonparametric Tests** from the datasheet **Analyze** menu to display the Specify *t*-test Analysis Window. This window displays a "list of variables in use" in your datasheet and the following types of *t*-tests:

- | | |
|---------------------------------|---|
| Two-group <i>t</i>-tests | Two-group tests require one Outcome variable and one Grouping variable that divide the sample into two groups. |
| Paired <i>t</i>-tests | Paired <i>t</i> -tests require two related Outcome variables such as weight before dieting, and weight after dieting. |

One-Group *t*-tests One-group tests require only one outcome variable.

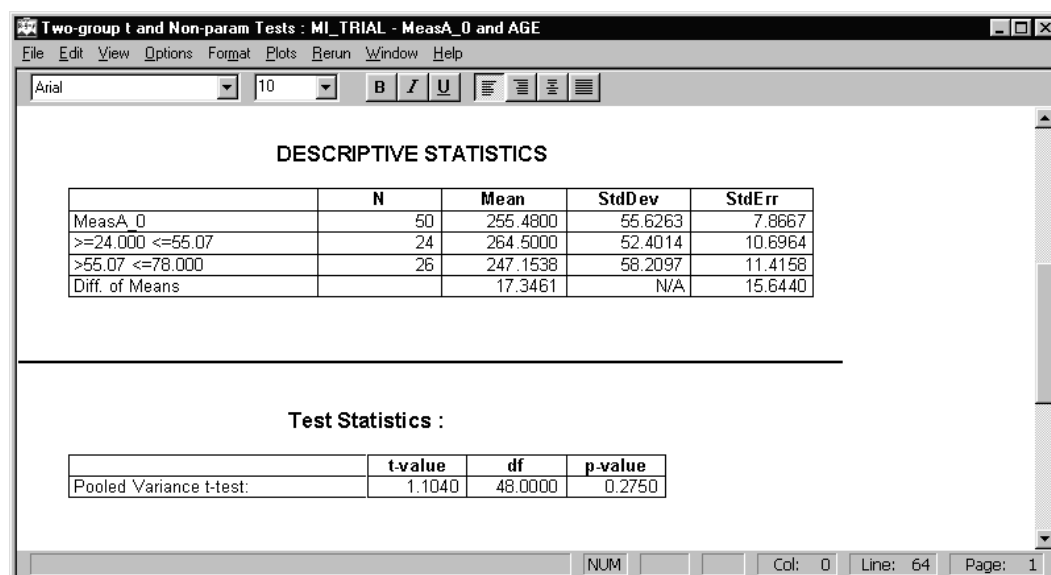


The 'Specify t-Test Analysis' dialog box is shown. It has three main sections: 'Two-group', 'Paired', and 'One-group'. The 'One-group' section is active. The 'Variables' list on the left contains: OBS, SYMPDUR, AGE, MeasA_0, MeasA_1, MeasA_2, MeasA_3, MeasB_0, and MeasB_1. The 'Outcome var.' field is empty. The 'Grouping var.' field is empty. The 'Null hypothesis : mean diff. =' field is set to 0.00. The 'OK', 'Cancel', 'Clear', and 'Help' buttons are at the bottom.

Drag and drop variable(s) from the Variables scrolled listbox into the requisite datafield for the type of *t*-test required (Two-group, Paired, or One-group). After the variable(s) have been placed in the datafield(s), the remaining *t*-tests selections are grayed out.

Two types of variables can be selected: Outcome (or Dependent) variables, and Grouping variables.

After pressing the **OK** button, the Output window showing the default results is displayed:



The output window shows the results of a Two-group t-test. The title bar is 'Two-group t and Non-param Tests : MI_TRIAL - MeasA_0 and AGE'. The menu bar includes File, Edit, View, Options, Format, Plots, Run, Window, and Help. The font is Arial, size 10. The output is divided into two sections: 'DESCRIPTIVE STATISTICS' and 'Test Statistics :'. The 'DESCRIPTIVE STATISTICS' section shows a table with columns N, Mean, StdDev, and StdErr. The 'Test Statistics :' section shows a table with columns t-value, df, and p-value.

	N	Mean	StdDev	StdErr
MeasA_0	50	255.4800	55.6263	7.8667
>=24.000 <=55.07	24	264.5000	52.4014	10.6964
>55.07 <=78.000	26	247.1538	58.2097	11.4158
Diff. of Means		17.3461	N/A	15.6440

	t-value	df	p-value
Pooled Variance t-test:	1.1040	48.0000	0.2750

Ungrouped Variables as Grouping Variables

If a variable is selected as a grouping variable, and it is ungrouped, you are warned that the variable does not have any grouping information, and you will be given an option to group values. If a grouping variable has more than two groups, the system asks you if you want to switch to ANOVA or re-group the categories.

When you choose to re-group, the Set Cutpoints /Set Categories window is displayed, and you can create two groups using one of four methods.

After choosing a method, and obtaining two groups, click on **OK**, the system displays the Specify *t*-test window, and the grouping variable is displayed in the appropriate field.

Hypotheses in *t*-tests

All hypothesis tests are two-sided tests, so the output includes two-tailed *p*-values. Refer to Two-group *t*-test, Paired *t*-test, or One-group *t*-test in the following sections.

Two-group *t*-test

The Two-group *t*-test tests for a specified difference between the means from two independent populations, using two independent samples that include an appropriate outcome variable. The null and alternative hypotheses are:

$$H_0 : \mu_1 - \mu_2 = c \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq c$$

where *c* is some value.

Usually, the mean difference tested is zero (i.e., *c* = 0). Sample sizes need not be equal. The two samples may be from an experiment in which cases are randomly assigned to one or two treatments, or they may arise from a survey in which some characteristic, such as gender, divides cases into two groups.

The Two-group *t*-test accepts an outcome variable and a grouping variable. You can use the Clear button to remove variables from the fields. The response, or outcome variable is the variable being used in the test, and the grouping variable is the variable that identifies the group membership of each case. The grouping variable is a binary variable of either the nominal, or ordinal type. If you try to use a continuous variable, or a variable with more than two groups, you will be warned to set cutpoints.

The usual null hypothesis being tested is that of equal population means (the difference in the population means is zero). If the sample means are far apart relative to the standard error of the mean difference, then you reject the null hypothesis and conclude that there is sufficient evidence that the population means are not equal. A small *p*-value gives less chance of an error in rejecting the null hypothesis. Sometimes the null hypothesis being tested is that the difference between the population means is a nonzero value. For example, the difference in population mean weight between cases following a certain diet, and those not on a diet, could be hypothesized to be 5 kilograms.

If an outcome variable is normally distributed, with equal variances in both populations, then the assumptions for making this test are met. If variances are not equal, the system provides a *t*-test using unequal variances. The *t*-test is quite robust to lack of normality unless the distribution is quite far from normal.

Sometimes you can largely overcome lack of normality, or unequal variances by the use of transformations, or by the use of robust statistics such as trimmed means and Winsorized

variances. Non-parametric tests can be used when you cannot assume a normal distribution for the variable of interest, but it still has a continuous distribution.

Paired *t*-test

The Paired section of the *t*-test window allows you to specify the outcome variables used in a paired or dependent sample *t*-test. Such samples are often the result of measuring the same variable at two different time periods, with or without a treatment intervention.

Alternatively a case in sample 1 may be paired or matched with a case in sample 2. This may be a natural pairing, as with twins, or an artificial pairing when an investigator matches two cases with similar characteristics thought to be related to the outcome variable. By the nature of the design, sample sizes of the two groups will be equal. Two outcome variables are needed, but no grouping variable is required.

The usual null hypothesis is that the difference in the population means is zero, but you can use a nonzero value:

$$H_0 : \mu_d = c \quad \text{versus} \quad H_0 : \mu_d \neq c$$

where μ_d is the population mean difference, and c is some value.

The differences between each pair are computed, as well as the mean and standard deviation of these differences. A paired *t*-test statistic is then computed by dividing the mean difference by the standard error of the mean difference (when the null hypothesis is zero mean difference). If differences in the sample means are large relative to the standard error of the mean difference, then the null hypothesis is rejected, and the conclusion is that the population means are not equal. The assumption made in using this *t*-test is that the differences are normally distributed. The test is quite robust, but if the assumption cannot be at least approximately met, then you can consider transformations on the differences, or the use of robust statistics such as the trimmed mean. The system also includes non-parametric options that assume a continuous distribution, but not normality.

One-group *t*-test

You can use a one-group test when you want to test that a population mean is a specified quantity. That quantity may be one that occurred in the past, or one that is expected by theory and given by:

$$H_0 : \mu = c \quad \text{versus} \quad H_0 : \mu \neq c$$

where c is some value.

The hypothesized value of the mean is usually not zero, but instead depends on the variable that is chosen. The paired *t*-test is a special case of a one-group *t*-test.

A single sample is assumed where the outcome variable being measured is assumed to follow a normal distribution. If the data are not normally distributed, then you should try transformations to get the distribution as close to normal as possible, or use a non-parametric test.

The *t*-test statistic is computed by dividing the difference between the sample mean and the hypothesized population mean, by the standard error of the sample mean. A large *t*-value or a small *p*-value signify that it is unlikely that the null hypothesis is true.

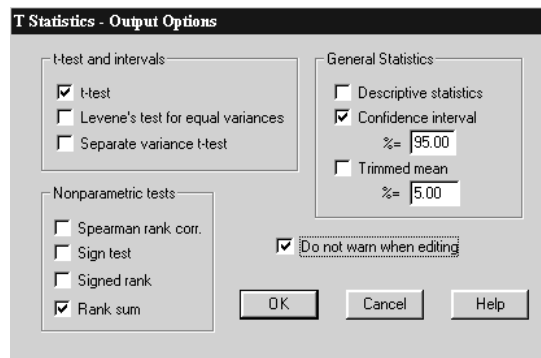
Output Options

From the *t*-test Output window **Options** menu select **Output** to display the Output Options window. The Output Options comprise three groups:

- ***t*-test and intervals:** *t*-test, Levene's test for equal variances, Separate variance test.
- **Nonparametric tests:** Spearman rank corr., Sign test, Signed rank, Rank sum.
- **General Statistics:** Descriptive Statistics, Confidence interval, Trimmed mean.

The **Levene's test** and **Separate variance *t*-tests** apply only to the Two Independent Group test. The **Descriptive statistics**, **Trimmed mean** options are available for all *t*-tests. The **Rank sum test** is available for the Two Independent Groups analysis.

The **Spearman rank correlation**, the **Sign test**, and the **Signed rank test** are available for the Paired tests. The **Sign test** and the **Signed rank test** are also available for the Single-Group tests.



Two-group Output Options

Dragging and dropping variables into the Outcome and Grouping variable datafields in the Two-Group area of the Specify *t*-test window, and then pressing the **OK** button, displays the Output window.

NOTE: You can view additional output by selecting **Output** from the **Options** menu in an Output window, selecting an option, and pressing the **OK** button.

Descriptive statistics

Default output includes the names of the two groups, sample size, mean, standard deviation, and standard error of the mean for each of the two groups. Standard error of the mean is the standard deviation divided by the square root of sample size of each group. When you request Trimmed mean, trimmed means and Winsorized standard deviations are also shown.

Confidence limits

You can request confidence limits in the range: greater than 50% and less than 100% for each of the two means. The default is 95%.

A 95% confidence limit has a 95% chance of covering the true population mean. In interpreting this interval, you are assuming that the data are normally distributed.

***t*-test for Two Independent Groups**

Output includes results from the *t*-test when variance estimates are pooled. A single pooled variance is a weighted average of the sample variances of the two samples. Weights used are sample sizes minus 1, or the degrees of freedom for each sample, respectively:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The two - sample test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - c}{s_p \sqrt{(1/n_1) + (1/n_2)}}$$

and is distributed as a Student's *t*-distribution with degrees of freedom equal to the sum of the sample sizes for the two groups minus 2 ($df = n_1 + n_2 - 2$).

The numerical value of the test statistic, its degrees of freedom, and associated *p*-value appear in the output. A small *p*-value signifies that the difference in population means is significantly different from the hypothesized value, *c*. If population variances are unequal, the stated *p*-value may be either too large or too small; a separate variance *t*-test is recommended.

Levene's test for equal variance

The results of the Levene's test for equal variance are displayed under Test Statistics. Levene's test computes the absolute deviation of each observation from its group mean, and then performs a two-sample *t*-test on the absolute deviations.

Output is based on an *F*-value from ANOVA (which is equivalent to the square of the *t*-value obtained from *t*-test). A large *F*-value, or a small *p*-value, is an indication of unequal variances. Levene's test has been shown to be quite robust for lack of normality, but it may not perform well for small sample sizes.

If the null hypothesis of equal variances is rejected, then you might consider using the Separate variance *t*-test that is given as an option. Click on the *t*- and Non-parametric Tests output window **Options** menu, and choose the Separate variance *t*-test. Alternatively, this may be an indication of lack of normality or outliers, so you can consider transformations, removal of outliers, and the Trimmed mean or Non-parametric tests options.

Separate variance *t*-test

The Separate variance *t*-test does not assume equal population variances when testing the same null hypothesis as pooled *t*-test. Thus, it does not use a pooled estimate of the population variance. The estimate of the variance of the mean difference is the sum of the variances of each sample mean.

The separate variance two-sample *t*-test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - c}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}}$$

and is approximately distributed as a Student's *t* distribution with degrees of freedom approximated by:

$$df = [\{k^2 / (n_1 - 1)\} + \{(1 - k)^2 / (n_2 - 1)\}]^{-1}$$

where:

$$k = \frac{s_1^2 / n_1}{(s_1^2 / n_1) + (s_2^2 / n_2)}$$

The approximate degrees of freedom may not be an integer (see Welch (1947)). A large absolute value of *t*, or a small *p*-value, indicates that the null hypothesis of equal means (the default) should be rejected.

The standard *t*-test assumes equal variances in the two groups and should be used if the variances are equal. The separate variance *t*-test should be considered if the Levene's test for equal variances rejects the null hypothesis of equal variances. Other alternatives are: transformations, data screening to remove possible outliers, robust statistics, or non-parametric tests.

Rank sum

The Rank sum or Mann-Whitney U test is available for two independent groups. This test yields results equivalent to the Wilcoxon rank sum test, and is called the Mann-Whitney - Wilcoxon test in some texts. Only independent and continuous distributions need to be assumed in making this test. Distributions of two populations are identical in form and differ only in location, rank sum has been shown to be useful in identifying differences in medians or means.

This test requires only about 5% more observations to have the same power as the *t*-test if the data are normally distributed. Note also that if you want to test that the difference in location is some nonzero value, you must add that hypothesized value to each of the observations in the first group before performing this non-parametric test.

The null hypothesis being tested is that the two samples come from identical populations. When the Observations from the two groups are combined and ranked from 1 to *N* with average ranks assigned to tied observations. The Mann-Whitney U statistic is reported:

rank sum of first group *R*₁ minus *n*₁ (*n*₁-1)/2 along with its significance level based on normal approximation. The test statistic is defined as:

$$\frac{R_1 - n_1(N+1)/2}{\sqrt{[n_1 n_2 (N+1) / 12] - [n_1 n_2 \sum (t_s^3 - t_s) / \{12N(N-1)\}]}}$$

where *t_s* is the number of tied observations.

Trimmed mean

A similar pooled variance *t*-test is reported when the Trimmed mean is requested. You should use a trimmed mean when you expect that outliers, or observations from a population other than the one of interest may be in the sample, and you wish to minimize the effect of these extraneous cases.

Each sample is trimmed by ordering cases from smallest to largest and then symmetrically omitting a percentage of the cases that are the extreme ones on both sides of the distribution. The default level of trimming is 5%, but you can specify your own level of trimming in the range of 0-25%.

Sample group standard deviations are computed using the Winsorizing method. This replaces the extreme values that are discarded in trimming by appropriate values that are from the sample but are less extreme. See Winsorized Standard Deviation under Descriptives (and Dixon And Massey, 1983). The test statistic, degrees of freedom, and *p*-value are given. A large *t* or a small *p*-value indicates that the null hypothesis should be rejected. The properties of this test statistic were investigated by Yuen and Dixon in 1973.

Paired Group Output Options

Dragging and dropping variables into the Variable1 and Variable 2 datafields in the Paired area of the Specify *t*-test window, and then pressing the **OK** button, displays the Paired *t* and Non-parametric test output window showing the output for a *t*-test comparing two paired or dependent samples.

NOTE: You can view additional output by selecting **Output** from the **Options** menu in an Output window, selecting an option, and pressing the **OK** button.

Descriptive Statistics

Output includes names of the groups, sample size, means, standard deviation, and standard errors of the means for each of the two groups and their difference. Standard error of the mean is standard deviation divided by the square root of sample size of each group. When you request trimmed means, you also get Winsorized standard deviations.

Confidence Limits

You can request greater than 0% and less than 100% confidence limits for each of the means. The default is 95%. A 95% confidence limit has a 95% chance of covering the true population mean. In interpreting this interval, you are assuming that the data from each population are normally distributed.

t-test for Paired Groups

The usual null hypothesis is that the difference in the population means is zero but you can supply other values. The system computes the difference between each pair of cases as well as the mean and standard deviation of these differences. A paired *t*-test is computed by dividing the mean difference by the standard deviation of the mean difference (when the null hypothesis is zero mean difference), and is given by:

$$t = \frac{\bar{d} - c}{\sqrt{S_d^2 / N}}$$

where \bar{d} is the mean, and S_d^2 is the variance of the difference in the paired X s, and $n_1=n_2=N$. The test statistic is distributed as a Student's t with $(N-1)$ degrees of freedom. If the differences in the sample means are large relative to the standard deviation of the means differences, then the default null hypothesis ($c=0$) is rejected. The conclusion is that the population means are not equal. Equivalently, the smaller the p -value, the smaller the likelihood that you make an error in rejecting the null hypothesis.

The assumption made in using this t -test is that the differences are normally distributed. The test is quite robust. However if the assumption cannot be at least approximately met, then you can consider transformations on the differences, or the use of robust statistics such as trimmed mean. The system also includes non-parametric options that assume a continuous distribution but not normality.

Trimmed t -test

A t -test is reported where trimmed means are used. You use a trimmed mean when you expect that outliers, or observations from a population other than the one of interest may be in the sample, and you want to minimize the effect of those extraneous cases. The sample is trimmed by ordering the cases from smallest to largest, and then symmetrically omitting a percentage of the cases that are the extreme ones on both sides of the distribution. The default level of trimming is 5%, but you can specify your own level of trimming.

The sample standard deviation is computed using the Winsorizing method. This replaces the extreme values that are discarded in trimming by appropriate values that are from the sample, but are less extreme (see Dixon and Massey, 1983).

For a matched pairs trimmed t -test, trimming and Winsorizing are performed on the paired differences rather than on the two variables.

The t statistic, degrees of freedom, and p -value (two-tailed test) are given. A large t or a small p -value indicates that the null hypothesis should be rejected.

Spearman rank correlation

Spearman rank correlation is essentially the product moment correlation between ranked data. Hence, it is the rank correlation of the cases in the first outcome variable versus those in the second outcome variable. In most textbooks, this formula is:

$$r_s = \frac{1 - 6D}{N^3 - N}$$

where D is the sum of squared differences in the ranks of the paired observations. The denominator applies only to the $6D$ and not the 1.

When ties occur, it is modified to:

$$r_s = \frac{A_1 + A_2 - D}{2\sqrt{A_1 A_2}}$$

where:

$$A_i = \frac{N^3 - N - T_i}{12}, i = 1, 2$$

$$T_i = \sum_j (t_{ij}^3 - t_{ij})$$

and t_{ij} is the number of observations tied with a single value for variable i . When N is at least 10, the Spearman rank correlation coefficient may be tested by a t -test based on $N-2$ degrees of freedom. The test statistic used is:

$$t = r_s \sqrt{(N-2) / (1-r_s^2)}$$

Sign Test (Matched)

The sign test does not use ranks. Instead the differences between the first and second paired variables are computed and are replaced by + and - signs (+ being used when the first is larger than the second). A binomial test is made and the null hypothesis being tested is that the probability of a + sign is 1/2.

Let N_+ and N_- be the number of positive and negative differences, N_{\min} be the minimum of N_+ and N_- , and $N_T = N_+ + N_-$ be the total number of nonzero differences. When N_T is at most 100, the two-tailed p -value is exactly calculated as:

$$pval = (1/2)^{(N_T-1)} \sum_{j=0}^{N_{\min}} \frac{N_T!}{j!(N_T-j)!}$$

where:

$$y! = \prod_{i=1}^y i = 1 \times 2 \times \dots \times y \text{ (factorial) and } 0! = 1.$$

For large samples, normal approximation is used to calculate the p -value.

Signed Rank (Matched)

The Wilcoxon signed rank test is computed by subtracting the values of the second variable from those of the first. The absolute value of this difference is obtained, along with its associated sign.

The sum of the ranks associated with the positive differences (R_+) are computed. The same is done for the negative ranks (R_-). The lesser of these two quantities (R_{\min}) is used to compute a statistic from which a p -value is derived based on a normal approximation:

$$z = \frac{[R_{\min} - N_T(N_T + 1)/4]}{\sqrt{N_T(N_T + 1)(2N_T + 1)/24}}$$

where N_T is the total number of nonzero differences. Exact probabilities are calculated when there are at most eight nonzero differences, or where the smaller of the ranks is at most 2.5.

The null hypothesis tested is that the median of the population is a specified value. In making this test, we are assuming that the population is continuous and symmetric.

Single Group Output Options

Dragging and dropping a variable into the Variable datafield in the One-group area of the Specify *t*-test window, and then pressing the **OK** button, displays the One-group *t* and Non-parametric test output window showing the output for a one-group comparison *t*-test.

NOTE: You can view additional output by selecting **Output** from the **Options** menu in an Output window, selecting an option, and pressing the **OK** button.

Descriptive Statistics

The output display includes variable names, sample size, mean, standard deviation, and standard error of the mean. The standard error of the mean is standard deviation divided by the square root of sample size. When you request trimmed means, you get trimmed mean and Winsorized standard deviations.

Confidence Limits

You can request greater than 50% and less than 100% confidence limits for each of the means. The default is 95%. A 95% confidence limit has a 95% chance of covering the true population mean. In interpreting this interval, you are assuming that the data from each population are normally distributed.

t-test for One Group

Results from the *t*-test include the *t* statistic, the degrees of freedom and the p -value. One group tests are used when you want to test whether the population mean is a specified quantity. This quantity may be one that occurred in the past or one that is expected by theory. The hypothesized value of the mean is usually not zero, but instead depends on the variable chosen.

single sample is assumed. The outcome variable being measured is assumed to follow a normal distribution. If the data are not normally distributed, then you should try transformations to get the distribution as close to normal as possible, or use a non-parametric test.

The *t*-test is computed by dividing the difference between the sample mean and the hypothesized population mean by the standard deviation of the sample mean:

$$t = \frac{\bar{x} - c}{\sqrt{s^2 / N}}.$$

The test statistic is distributed as a Student's *t* with *N* - 1 degrees of freedom. A large *t* value or a small *p*-value signifies that it is unlikely that the null hypothesis is true.

Trimmed Mean Test

A *t*-test is displayed where a trimmed mean is used. A trimmed mean is used when you expect that outliers, or observations from a population other than the one of interest may be in the sample, and you wish to minimize the effect of these extraneous cases. The sample is trimmed by ordering the cases from smallest to largest, and then symmetrically omitting a percentage of the cases that are the extreme ones on both sides of the distribution. The default level of trimming is 5%, and you can also supply your own trimming level.

The sample standard deviation is computed using the Winsorized method. Winsorizing replaces the extreme values that are discarded in trimming by appropriate values that are from the sample but are less extreme (see Dixon and Massey, 1983).

The *t* statistic, degrees of freedom, and *p*-value (two tailed test) are given. A large *t* or a small *p*-value indicates that the null hypothesis should be rejected.

Sign Test for One Group

The sign test is based on the sign of the difference between the chosen variable and the hypothesized median. The level of significance reported is for a two-sided test in which the null hypothesis is that of equal numbers of + and - signs above and below the median. You assume that the cases are independent and continuous.

Let N_+ and N_- be the number of positive and negative differences, N_{\min} be the minimum of N_+ and N_- , and $N_T = N_+ + N_-$ be the total number of nonzero differences. When N_T is at most 100, the two-tailed *p*-value is exactly calculated as:

$$p \text{ value} = (1/2)^{(N_T-1)} \sum_{j=0}^{N_{\min}} \frac{N_T!}{j!(N_T-j)!},$$

where:

$$y! = \prod_{i=1}^y i = 1 \times 2 \times \dots \times y \text{ (factorial) and } 0! = 1.$$

For large samples, normal approximation is used to calculate the *p*-value.

Signed Rank Test

The Wilcoxon signed rank test is computed by subtracting the values of the second variable from those of the first. The absolute value of this difference is obtained, along with its associated sign. The sum of the ranks associated with the positive differences (R_+) are computed. The same is done for the negative ranks (R_-). The lesser of these two quantities (R_{\min}) is used to compute a statistic from which a p -value is derived based on a normal approximation:

$$z = \frac{[R_{\min} - N_T(N_T + 1)/4]}{\sqrt{N_T(N_T + 1)(2N_T + 1)/24}}$$

Where N_T is the total number of nonzero differences. Exact probabilities are calculated when there are at most eight nonzero differences, or where the smaller of the ranks is at most 2.5. The null hypothesis tested is that the median of the population is a specified value. In making this test we are assuming that the population is continuous and symmetric.

Plots for *t*-test

The system provides four types of plots for the visual check of assumption used in *t*-test.

Scatterplot

A Scatterplot of the first variable against the second variable is available for paired *t*-test. You can use this to assess any association between two variables. For more information on this plot, see the discussion under Plots later in this manual.

Boxplot

A Boxplot is available for all types of *t*-test. For the two-group *t*-test, a Boxplot is drawn for each group. For a paired *t*-test, three Boxplots are drawn, one for each variable, and one for the differences. For the one-group *t*-test, the system displays a Boxplot corresponding to the outcome variable. For more information on this plot, see the discussion under Plots later in this manual.

Histogram

A Histogram is available for all types of *t*-test and the histogram display is similar to the Boxplots described above. For more information on this plot, see the discussion under Plots later in this manual.

Means Comparison Chart

A Means Comparison Chart is available for the two-group *t*-test. For more information on this plot, see the discussion under Plots later in this manual.

6. Analysis of Variance (ANOVA)

USING ONE-WAY ANOVA

USING TWO-WAY ANOVA

OUTPUT OPTIONS

PLOTS FOR ONE-WAY AND TWO-WAY ANOVA

General

Analysis of Variance (ANOVA) is a statistical method for analyzing differences between means of sets of samples. The sets are differentiated by a factor whose influence on the means of the groups is to be investigated.

The Specify ANOVA window allows you to scroll through a list of the variables being used from your datasheet, and choose a One-way, or a Two-way ANOVA.

One-way ANOVA is used when you have one factor grouped into two or more samples, and you want to test whether or not the population means are equal. The independent samples may be outcomes of different experimental treatments, or derived from a survey and, based on some factor such as religion, grouped.

Two-way ANOVA is used when you have two factors (each grouped), and you want to simultaneously test their effects on the means. To study the means, it is necessary to analyze their variances, so the tests have been called analysis of variance or ANOVA. Users unfamiliar with ANOVA should read a statistical text that discussed this topic, such as Dixon and Massey (1983) or Dunn and Clark (1987).

One-way ANOVA Model

A one-way ANOVA model may be written as:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, p \text{ and } j = 1, \dots, n_i,$$

where:

y_{ij} is the j th observation of outcome variable Y belonging to the i th population;

μ_i is the i th population mean of Y ;

ε_{ij} s are error terms which are independently and identically distributed as normal mean zero and constant variance σ_ε^2 (i.e., $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$);

n_i is the number of samples from the i th population, $N = \sum_{i=1}^p n_i$; and:

p is the number of populations being compared.

The population means are parameterized as $\mu_i = \mu + \alpha_i$

Where:

μ_i is the overall mean; and α_i is the i th main effect, $\sum_{i=1}^p \alpha_i = 0$.

A consequence of the one-way model above is that $y_{ij} \sim N(\mu_i, \sigma_\epsilon^2)$

The different populations may have differing means, but all have the same variance, σ_ϵ^2 .

Estimates of population means are given by sample group means, as in:

$$\hat{\mu}_i = \bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij} / n_i.$$

An estimate of the variance, σ_ϵ^2 , is derived from a weighted average of within-group variances.

This estimate is called the pooled estimate of variance.

Hypothesis

In one-way ANOVA, the null hypothesis being tested is of equal population means:

$$H_0: \mu_1 = \dots = \mu_p \text{ vs } H_1: \text{at least two means are different}$$

If the various sample group means are far apart, then the null hypothesis will likely be rejected. The test used is an F test. The larger the F -value in the F -value column of the ANOVA Table, or the smaller the p -value, the less chance you have of making an error in rejecting the null hypothesis.

Rejecting the null hypothesis of equal means does not tell you which means are different from each other. Further testing using contrasts on the means may be necessary.

Two-way ANOVA Model

A two-way ANOVA model may be defined as:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, i = 1, \dots, p, j = 1, \dots, q, k = 1, \dots, n_{ij},$$

where:

y_{ijk} is the k th observation from a sub-population defined by the i th level of the first factor and the j th level of the second factor:

μ_{ij} is a population mean;

ϵ_{ijk} s are error terms which are independently and identically distributed as normal mean zero and constant variance σ_ϵ^2 (i.e., $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$);

n_{ij} is the sample size of the sub-sample of outcomes drawn from the sub-population defined by the i th level of the first factor, and the j th level of the second factor:

$$\bar{y}_{ij.} = \sum_{k=1}^{n_{ij}} y_{ijk} / n_{ij}, ; \text{ and:}$$

p is the number of levels of the first factor, and q is the number of levels of the second factor.

The mean μ_{ij} is further decomposed into two main effects due to the two factors and an interaction effect between the two factors:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

where:

μ is an overall mean

α_i is the i th main effect of the first factor, $\sum_{i=1}^p \alpha_i = 0$;

β_j is the j th main effect of the second factor, $\sum_{j=1}^q \beta_j = 0$;

γ_{ij} is an interaction effect $\sum_{i=1}^p \gamma_{ij} = 0$; and $\sum_{j=1}^q \gamma_{ij} = 0$;

A consequence of the full two-way model (i.e., model with interaction terms) is that:

$$y_{ijk} \sim N(\mu_{ij}, \sigma_{\epsilon}^2).$$

The different populations may have differing means, but all must have the same variance, σ_{ϵ}^2 . The full two-way ANOVA model may be expressed as a one-way ANOVA in which the levels of the only factor are derived from the combination of the levels of the two factors in two-way ANOVA. Estimates of population means are given by sample group means, as in

$$\hat{\mu}_{ij} = \bar{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk} / n_{ij}.$$

An estimate of the variance, σ_{ϵ}^2 , is derived by a weighted average of within-group variances. This estimate is called the pooled estimate of variance.

Most texts on ANOVA displays data for two-way ANOVA in a two-way table where rows correspond to levels of the first factor (e.g., male and female) and columns represent levels of the second factor (e.g., married, never married, divorced or separated, and widowed). The interior of this two by four table consists of eight cells with observations placed in the appropriate cell. Means are calculated for each cell (defined by the row by column combination). Marginal (row or column) and overall means may also be computed.

With the system, you can specify that you want to omit an interaction term. In an additive model, the mean is represented as :

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

If you have only a single observation in each combination of the two factors, an additive model is the only model which allows you to perform hypothesis tests. Otherwise, you will have a saturated model and your model will fit your data perfectly.

Hypotheses

The overall null hypothesis being tested is the hypothesis of equal population means defined by two factors:

$$H_0 : \mu_{ij} = \mu \text{ for all } i \text{ and } j \quad \text{versus} \quad H_1 : \text{at least two means are different.}$$

This overall hypothesis can be broken down into three hypothesis tests on main effects and interaction effects:

$H_0: \alpha_i = 0$ for all i versus $H_1: \alpha_i \neq 0$ for some i ;

$H_0: \beta_j = 0$ for all j versus $H_1: \beta_j \neq 0$ for some j ; and

$H_0: \gamma_{ij} = 0$ for all i and j versus $H_1: \gamma_{ij} \neq 0$ for some i and j .

The first two are significance tests on main effects of the two factors. The third is a test of significance on the interaction effect. The third does not apply when there is only one observation per cell, or when an additive model is requested. Zero interaction will exist if, for example, the effects of gender and marital status are strictly additive.

If graduating from high school adds \$5000 to the average yearly income of both females and males, then gender and high school graduation status are additive. If, on the average, males get \$10,000 more and females \$3000 more for graduating from high school, then the effects of gender and graduation status are not additive on average yearly income.

ANOVA uses an F -test. The larger the F -value in the F -value column of the ANOVA Table, or the smaller the p -value, the less chance of an error in rejecting the null hypothesis.

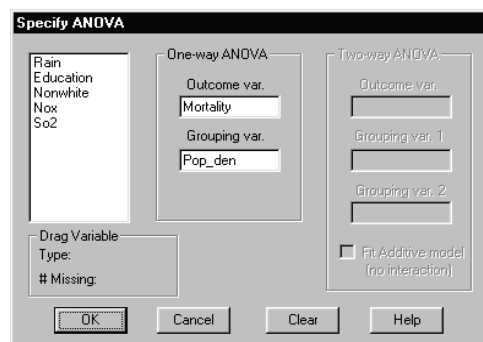
Using One-way ANOVA

One-way ANOVA is used to test the effects of one factor grouped into two or more samples.

1. Select the **ANOVA** option from the datasheet **Analyze** menu to display the Specify ANOVA window.
2. Drag and drop the desired variable from the Variables listbox into the Outcome var. datafield.
3. Drag and drop a grouping variable into the Grouping var. datafield.

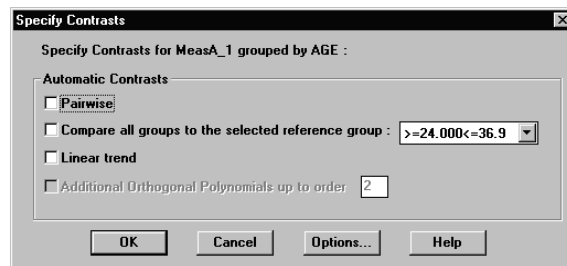
NOTE: For example, the factor could be gender (male or female), or could be marital status (married, never married, divorced or separated, and widowed).

4. If the selected grouping variable is a continuous variable, you will be asked to group those variables.
5. Click the **OK** button to display the Output window.



Specify Contrasts

To specify contrasts, click on the **Options** menu of the ANOVA Output window, and select **Contrasts** to display the Specify Contrasts window:



Four different types of contrasts can be selected:

Pairwise

This selection provides a pairwise comparison of each mean with every other mean..

Compare all groups to the selected reference group

This selection compares all sample means to a reference group sample mean. Descriptive statistics along with the pooled variance and separate variance *t*-tests are displayed

Linear trend and Orthogonal polynomials

If the grouping variable is continuous/ordinal, then testing for linear trend and/or orthogonal polynomials may be useful in deciding on the relationship between the grouping variable and the outcome or dependent variable. The methods of linear trend and orthogonal polynomials implicitly assume that the levels of the grouping variable are equally spaced. The maximum order of polynomials contrast is the number of levels minus 1. The polynomials will be orthogonal to each other, provided that the sample sizes in each group are all equal (i.e., you have a balanced design).

A large F-value for a linear contrast indicates that the linear trend is significant. Similar conclusions apply to orthogonal polynomial contrasts. The lowest value that you can enter is 2. The highest value is 1 less than the number of groups. The polynomials will be orthogonal to each other, provided that the sample sizes in each group are all equal (i.e., you have a balanced design). A large F-value for a linear contrast indicates that the linear trend is significant. Similar conclusions apply to orthogonal polynomial contrasts.

NOTE: You can include the Kruskal-Wallis nonparametric test in your ANOVA output (see *Output Options* later in this chapter). The test uses the ranks of the data, assuming continuous data. The null hypothesis being tested is that the *k* samples are drawn from *k* identical populations.

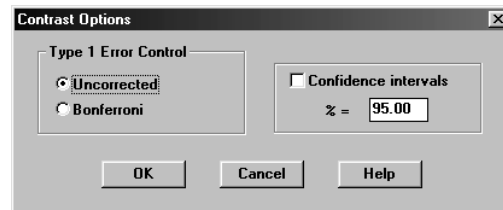
If you have unequal variances, we recommend considering transforming the data. This is particularly recommended when the sample sizes are quite different.

NOTE: The last two contrast selections (linear trend and orthogonal polynomials) require the grouping variable to be continuous/ordinal and equally spaced, since the grouping variable is being used as if it were an X variable in regression.

The first two contrast selections can be used for pairwise comparisons of the means; comparing one mean against another.

Contrast Options

Bonferroni adjustments and Confidence intervals may be specified using the **Options** button in the Specify Contrasts window



Variances – ANOVA One-way

If outcome variable Y is normally distributed with equal variances in the populations, then the assumptions for making the tests are met. If variances are not equal, then the system provides two F tests that allow for unequal variances (we recommend the Welch test). The Separate Variance output may be used for pairwise comparisons of means in such cases. However, for linear and orthogonal polynomial contrasts, equal variances are assumed.

Lack of equal variances in the groups, particularly if the magnitude of variance is associated with the size of the mean in the respective group, is often a sign of non-normality. Unequal variances may also occur if there are outliers. If your concern is outliers, you should screen your data for outliers, and remove them, or use the trimmed ANOVA. If your data are decidedly not normally distributed, you should consider transformations or use the Kruskal-Wallis nonparametric test.

Using Two-way ANOVA

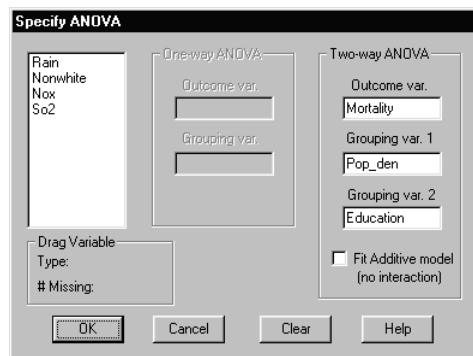
Two-way ANOVA is used to test the effects of two factors in combination.

1. Select the **ANOVA** option from the datasheet **Analyze** menu to display the Specify ANOVA window.
2. Drag and drop the desired variable from the Variables listbox into the Outcome var. datafield.
3. Drag and drop a grouping variable into the Grouping var. datafield.

NOTE: For example, one factor could be gender (male or female) and another could be marital status (married, never married, divorced or separated, and widowed).

4. If the selected grouping variable is a continuous variable, you will be asked to group those variables.

5. Click on the **OK** button to display the Output window.



Variances – ANOVA Two-way

If an outcome variable Y is normally distributed with equal variances in the populations, then the assumptions for making the tests are met. If variances are not equal, the system provides two F tests that allow for unequal variances (we recommend the Welch test). Lack of equal variances in the groups, particularly if the magnitude of variance is associated with the size of the mean in the respective group, is often a sign of non-normality. Unequal variances may also occur if there are outliers.

If your concern is outliers, you should screen your data for outliers and remove them, or use the trimmed ANOVA. If your data are decidedly not normally distributed, you should either consider transformations or use the Kruskal-Wallis nonparametric test.

Sometimes a large computed F (small p -value) will be obtained for the interaction term when the investigator thinks that the model should be additive. When this happens, it may be worthwhile to check for outliers, try to figure out if some other extraneous factor is present, or check for normality of the data. Transformations can sometimes be found to decrease the interaction effect. In general, it is simpler to interpret the results of two-way ANOVA if the interaction effect is not significant. It has been noticed that, in practice, significant interaction effects often occur when the row means and the column means are highly significantly different.

Output Options

The ANOVA output is displayed after pressing the **OK** button in a Specify ANOVA window.

Descriptive statistics and the ANOVA table appear first. For one-way ANOVA, there is only one outcome variable and one grouping variable or factor. You can request output similar to that for the two independent sample t -test. Here two or more samples or groups can be used. For two-way ANOVA, there are two required grouping variables, one for each factor.

NOTE: The system only allows complete cells (i.e., at least one observation falls into each of the cells defined by the factor levels).

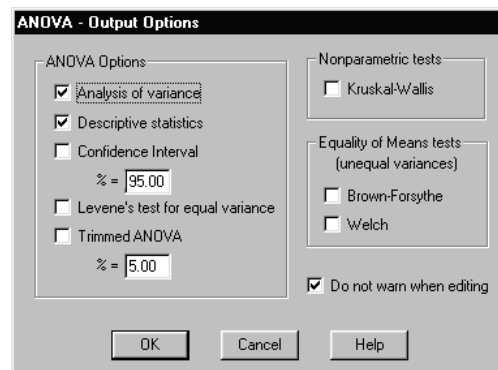
You can also modify the output options by choosing the **View** menu from the SOLAS 3.0 Main window and selecting **System Preferences**.

One-way ANOVA Output

The Output window for a one-way ANOVA, comparing two or more independent groups, initially displays the default output results. Additional output is available by choosing the **Options** menu and then selecting **Output** to display the Output window.

Default Output options for the one-way ANOVA include Analysis of Variance and descriptive statistics. Other options that are available are:

- Confidence limits, Levene's test for equal variance, and trimmed ANOVA.
- Nonparametric test (Kruskal-Wallis), Brown-Forsythe and Welch tests of Equality of Means.



Contrasts on population means may also be performed using the sample group means. If you reject the null hypotheses of equal population means, it may not be clear which means differ. To help determine which groups are different, click on the **Options** menu of the ANOVA output window, and select **Contrasts** (see Specify Contrasts earlier in this chapter).

Formulae

ANOVA table

The ANOVA table below provides standard results for the test of the null hypothesis of equal means, omitting only the p -value column:

Source	Sum of Squares	df	Mean Square	F-value
Grouping Variable	$SS_{treat} = \sum_{i=1}^p n_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$p-1$	$MS_{treat} = \frac{SS_{treat}}{(p-1)}$	$F = \frac{MS_{treat}}{MS_{error}}$
Error	$SS_{error} = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$	$N-p$	$MS_{error} = \frac{SS_{error}}{(N-p)}$	

Error Mean Square (or residual mean square) is a pooled estimate of the within-group variances; this is an estimate of the variance of the error terms, σ_e^2 . The so-called treatment or group Mean Square is displayed above the Error Mean Square. For example, if the grouping variable is

Education, this will be called the Education Mean Square. It is a weighted sum of the squared deviations of each sample mean from the overall mean, divided by the number of samples minus 1. The weights are the respective sizes in each group.

The *F*-test is computed by dividing the treatment mean square by the error mean square. A large *F*-value or a small *p*-value indicates that there is a small chance of making an error in rejecting the null hypothesis.

Descriptive Statistics

Default output includes names of the groups, sample size, mean, standard deviation, standard error of the mean, minimum, and maximum for each group.

Standard error of the mean is standard deviation divided by the square root of sample size of each group. When you request trimmed means, you also get robust standard deviations, and Winsorized standard deviations.

Robust standard deviation (labelled as Robust SD) is a robust estimate of the standard deviation based on mean absolute deviation. For the whole data of size *N*, it is defined as:

$$\text{Robust SD} = \sqrt{\frac{N\pi}{2(N-1)}} \left[\frac{\sum_{i=1}^N |y_i - \bar{y}|}{N} \right]$$

Where \bar{y} is the sample mean.

Confidence Intervals

You can request greater than 0% and less than 100% confidence limits for each of the means. The default is 95%. A 95% confidence limit has a 95% chance of covering the true population mean. In interpreting this interval, you are assuming that the data from each population are normally distributed.

Levene's Test for Equal Variances

Results of the Levene's test for equal variances appear after the ANOVA table output. Levene's test is done by computing the absolute deviation of each observation from its group mean, then performing a one-way ANOVA on the absolute deviations. A large *F*-value, or a small *p*-value is an indication of unequal variances. Levene's test has been shown to be quite robust for lack of normality, but it may not perform well for small sample sizes.

If the null hypothesis of equal variances is rejected, you might consider using the Welch or Brown-Forsythe tests. Alternatively, rejection of the null hypothesis may be an indication of lack of normality or outliers. Thus, you can consider transformations, removal of outliers, the trimmed means option or nonparametric tests.

Trimmed ANOVA

A trimmed ANOVA is used when you expect that outliers or observations from a population other than the one of interest may be in the sample, and you wish to minimize the effect of those extraneous cases. Each sample is trimmed by ordering cases from smallest to largest, then symmetrically omitting a percentage of the cases that are the extreme ones on both sides of the distribution. See Trimmed Mean under Descriptives. The default level of trimming is 5%, but you can specify your own level in the range 0-25%.

Sample group standard deviations are computed by Winsoring which replaces the extreme values that are discarded in trimming by appropriate values that are from the sample, but are less extreme. See Winsorized Standard Deviation under Descriptives (and Dixon and Massey, 1983).

A standard ANOVA is performed using trimmed means and Winsorized mean squares. A large F , or a small p -value indicates that the null hypothesis should be rejected. Properties of this test statistic for the two-sample cases were investigated by Yuen and Dixon in 1973.

Kruskal - Wallis Nonparametric Test

The Kruskal-Wallis test is available for one-way ANOVA. This Kruskal-Wallis test tests the null hypothesis of equal distributions in the samples using ranks. It assumes at least ordinal data.

Let N be classified into p groups with the i th group containing n_i cases. All cases from the p groups are combined, then ranked from 1 to N with tied cases assigned the average rank of the tied cases.

Let r_i be the sum of the ranks for the i th group. The Kruskal-Wallis statistic (KW) is defined as:

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^p (R_i^2 / n_i) - 3(N+1).$$

where ties occur, KW is modified to:

$$KW' = KW / \left[1 - \sum (t_s^3 - t_s) / (N^3 - N) \right]$$

where t_s is the number of observations tied with a single value, and summed over distinct values. If the minimum group size is greater than five, KW has a chi-square distribution with $(p-1)$ degrees of freedom. For small group sizes, see Table O in Siegel (1956.)

Brown-Forsythe Test

The Brown-Forsythe test is an approximate test that may be used when Levene's test rejects the null hypothesis of equal variances. The test statistic is defined as:

$$BF = \frac{\sum_{i=1}^p n_i (\bar{y}_i - \bar{y}_{..})^2}{\sum_{i=1}^p (1 - n_i / N) s_i^2}$$

where s_i^2 is the i th within-group sample variance:

$$s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (n_i - 1).$$

Critical values are obtained from an F distribution with $(p - 1)$ and df degrees of freedom. The denominator degrees of freedom df are implicitly defined by the Satterthwaite approximation:

$$df = \left[\sum_{i=1}^p c_i^2 / (n_i - 1) \right]^{-1},$$

where:

$$c_i = \frac{(1 - n_i / N) s_i^2}{\sum_{i=1}^p (1 - n_i / N) s_i^2}$$

See Brown and Forsythe (1974).

Welch Test

The Welch test is an approximate test that may be used when Levene's test rejects the null hypothesis of equal variances. The test statistic is defined as:

$$W = \frac{\sum_{i=1}^p w_i (\bar{y}_i - \tilde{y})^2 / (p - 1)}{1 + 2(p - 2) \sum_{i=1}^p \left[(1 - w_i / u)^2 (n_i - 1) \right] / (p^2 - 1)}$$

where:

$$w_i = n_i / s_i^2,$$

$$s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - y_i.)^2 / (n_i - 1),$$

$$u = \sum_{i=1}^p w_i$$

and:

$$\tilde{y} = \sum_{i=1}^p w_i \bar{y}_i / u.$$

When all population means are equal (even if the population variances are unequal), W has an approximate F distribution with $(p - 1)$ and df degrees of freedom. The denominator degrees of freedom df are defined as :

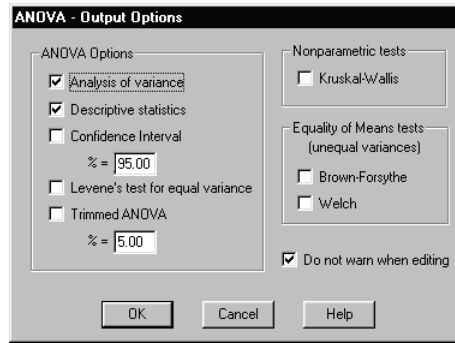
$$df = \left[3 \sum_{i=1}^p \{ (1 - w_i / u)^2 / (n_i - 1) \} / (p^2 - 1) \right]^{-1} \text{ See Welch (1947).}$$

Two-way ANOVA Output

The Output window for a Two-way ANOVA, comparing two or more independent groups, initially displays the default output results. Additional output is available by choosing the **Options** menu and then selecting **Output** to display the Output window.

Default Output options for the one-way ANOVA include Analysis of Variance and descriptive statistics. Other options that are available are:

- Confidence limits, Levene's test for equal variance, and trimmed ANOVA.
- Brown-Forsythe and Welch tests of Equality of Means.



The **Contrasts** option from the **Options** menu in the Output window is not available for two-way ANOVA, however, you **can** analyze two-way ANOVA as a one-way ANOVA.

ANOVA table

The ANOVA table below gives the standard results for the test of the null hypotheses: no interaction and equal means for both factors (omitting only the p-value column):

Source	Sum of Squares	df	Mean Square	F-Value
A	$SS_A = RSS(B, A * B) - RSS(A, B, A * B)$	$p-1$	$MS_A = \frac{SS_A}{(p-1)}$	$F_A = \frac{MS_A}{MS_{err}}$
B	$SS_B = RSS(A, A * B) - RSS(A, B, A * B)$	$q-1$	$MS_B = \frac{SS_B}{(q-1)}$	$F_B = \frac{MS_B}{MS_{err}}$
Interaction (A*B)	$SS_{A*B} = RSS(A, B) - RSS(A, B, A * B)$	$(p-1)(q-1)$	$MS_{A*B} = \frac{SS_{A*B}}{(p-1)(q-1)}$	$F_{A*B} = \frac{MS_{A*B}}{MS_{err}}$
Error	$SS_{err} = RSS(A, B, A * B)$ $= \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$	$N-pq$	$MS_{err} = \frac{SS_{err}}{(N-pq)}$	

The term $RSS(\bullet)$ denotes the residual sum of squares after a particular model is fit. The arguments inside the parentheses are the terms included in the model. For example, the error sum of squares (denoted by $RSS(A, B, A * B)$) is the residual sum of squares when a full two-way ANOVA model is fit:

$$RSS(A, B, A * B) = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk})^2, \text{ where: } \hat{y}_{ijk} = \bar{y}_{ij}.$$

The square root of the Error mean square (or residual mean square) is a pooled estimate of the standard deviation of the individual cases in the cells. The Interaction mean square and the mean square for each of the two factors, or grouping variables, are displayed above the Error mean square. An F test is performed by dividing each of the mean squares by the Error mean square, regardless of the magnitude of the F test for the interaction.

F_A is the test statistic used for testing the significance of factor A main effect; F_B is used for testing the significance of factor B main effect; and F_{A*B} is used for testing the significance of the interaction effect.

Each of these test statistics has an F distribution with the numerator degrees of freedom equal to the degrees of freedom associated with the particular effect and the denominator degrees of freedom equal to the Error degrees of freedom ($N-pq$). These three F tests will be independent of each other provided that you have equal sample sizes in each cell (i.e., you have a balanced design).

A large F -value or a small p -value indicates that there is a small chance of making an error if you reject the null hypothesis. Some analysts will fit an additive model after testing for significance of the interaction effect before testing for significance of the main effects (i.e., pool the Interaction sum of squares with the Error sum of squares). See Winer et al. (1991) for consequences of pooling strategies.

For the definitions of the following output options, see the one-way ANOVA output discussion earlier in this chapter:

Descriptive Statistics, Confidence Intervals, Levene's Test for Equal Variances, Trimmed ANOVA, Brown-Forsythe Test, and the Welch Test.

Plots for one-way and two-way ANOVA

By clicking on the **Plots** menu in the Output window, you can plot your data defined by groups. Five options are available for one-way ANOVA:

- Boxplots, Histograms, Means Comparison charts, Scatterplot of the group standard deviations versus the group means, and Box-Cox diagnostic plots.

For a two-way ANOVA, the groups are defined by the combination of levels of the two factors, or grouping variables. The following four plots are available:

- Boxplot, Histogram, Means Comparison charts, Scatterplot of the group standard deviations versus the group means.

Scatterplot of Group SDs vs Means

A scatterplot of the standard deviation of each group, versus the corresponding group mean, helps in assessing whether or not the assumption of equal variances in the group is met. If data are normally distributed, the sample mean and variances are independent. In practice, you will often find that the variance (or its square root - the standard deviation) increases as the mean increases. This can be taken as a convenient way of assessing normality in ANOVA, as the data are given by treatment groups. An increasing linear relationship between the standard deviation and the mean formation is appropriate (see Dunn and Clark, 1987).

Box-Cox Diagnostic Plot

The Box-Cox Diagnostic Plots option provides a log mean by log standard deviation plot for one- and two- way ANOVA. Clicking on the **Box-Cox Diagnostic Plots** option in the ANOVA **Plots** menu causes two trimming choices to be displayed; **Untrimmed** and **Trimmed Box-Cox**. Choosing either causes the plot to be displayed with its own statistics window.

The Box-Cox diagnostic plot is used not only to assess normality in grouped data, but also to suggest an appropriate transformation if data are not normally distributed. Two diagnostic plots are available.

When you perform an ANOVA with no trimming, the Box-Cox diagnostic plot will show the log of the means on the horizontal axis versus the log of the standard deviations on the vertical axis.

When you select the Trimmed ANOVA output option, the Box-Cox plot will show the log of the trimmed means versus the log of the winsorized standard deviations. The latter is useful if outliers are present.

The slope b of the regression line provides an estimate of the suggested power transformation of the outcome variable Y . The suggested transformation is

$$\text{Transformed } Y = Y^{(1-b)}.$$

See Box and Cox (1964).

For more information about plots see Chapter 7 of this manual.

7. Plots

SCATTERPLOT

HISTOGRAM

BOXPLOT

BAR CHART

MEANS COMPARISON CHART

NORMAL PROBABILITY PLOT

General

This chapter gives descriptions, and examples, of the plots available in the system, and that are most commonly used in statistical analysis. The menus in a plot output window make available a variety of functions that allow you modify an existing plot, or generate new plots using different data. Some of functions that you can use are:

- ◆ Perform transformations on your plot variables.
- ◆ Modify the scaling.
- ◆ Change point size of symbols.
- ◆ Set the plot axes and origin.
- ◆ Use points in local and global modes.
- ◆ Apply lines as linear fit, and mean x and y.
- ◆ Display statistics and diagnostics.

The Plot Output window menu functions are described in detail in Chapter 8 - Tutorial.

Scatterplot

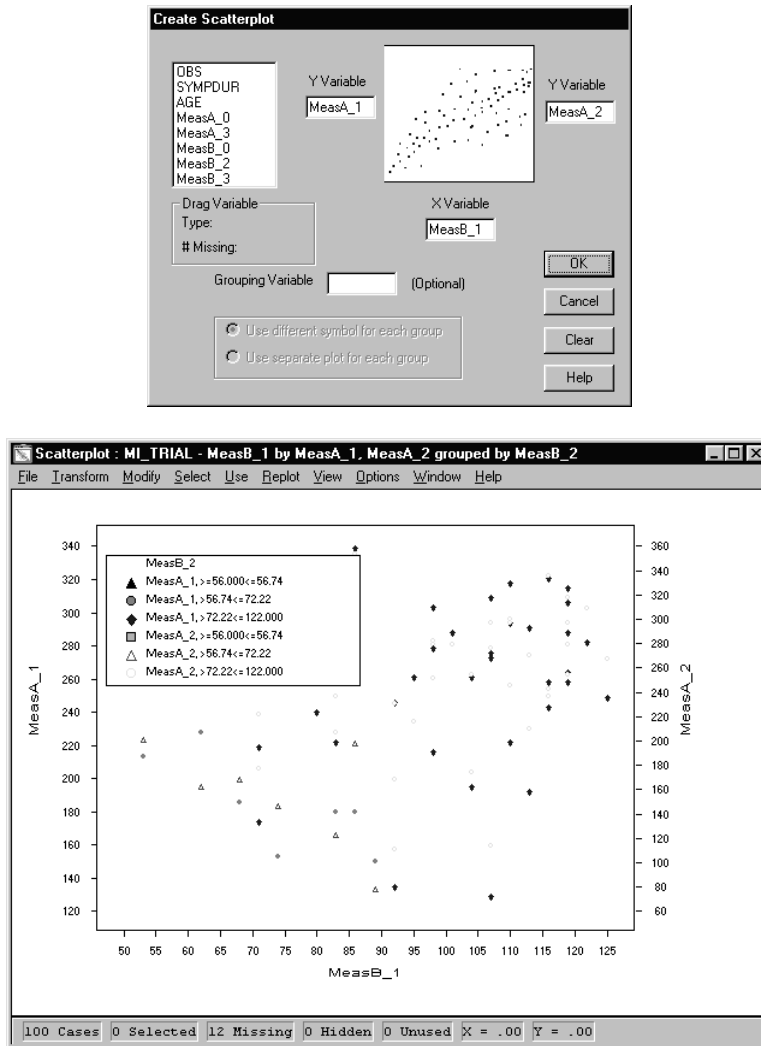
Scatterplot is the method used most often for displaying relationships between two ordinal or continuous variables. Usually an outcome or dependent variable is plotted on the vertical axis and an independent or predictor variable is plotted on the horizontal axis. It simultaneously shows all data for both variables as well as the interrelationship between the two variables. It is widely used in data screening and as a graphical diagnostic tool (e.g., normal probability plot is a scatterplot).

The system allows you to plot two variables along the vertical axis with two different tick mark labels on opposite sides of the plot. This feature is useful for overlaying two different plots with the same variable on the horizontal axis. In addition, the system has added reference lines including a regression line to aid you in interpreting your plot.

Creating a Scatterplot

Select **Scatterplot** from the Datasheet **Plot** menu, the system displays the Create Scatterplot window. The listboxes and datafields are described below:

Variables Listbox	Displays the list of used variables in the open datasheet.
Type	Displays the type of the variable currently selected.
Role	Displays the role of the variable currently selected.
Y Variable (left) datafield	Drag and drop a variable from the Variables listbox to the Y Variable datafield.
Y Variable (right) datafield	Drag and drop a variable from the Variable listbox to the Y Variable datafield. Adding a second Y variable lets you see two Y variables plotted against one X variable.
X Variable datafield	Drag and drop a variable from the Variable listbox to the X Variable datafield.
OK button	An X variable and a Y variable (left) must be specified before the OK button becomes available.
Grouping Variable datafield	If you enter a variable and it is not grouped, the system will prompt you to group the variable. Once a grouping variable is selected, other relevant options become available.
Use different symbol for each group checkbox	Shows all the groups in one plot, denoted by different symbols. This choice is grayed out unless you have chosen a grouping variable. The maximum number of groups is 60.
Use separate plot for each group checkbox	Shows an individual plot for each group. This choice is grayed out unless you have chosen a grouping variable. The maximum number of groups is 12.



Entering the required variables and pressing the **OK** button in the Create Scatterplot window displays an output window as shown above.

Histogram

Histogram is one of the most widely used methods of displaying univariate continuous data. It is used to display the shape of the distribution and is often compared with the symmetric bell-shaped normal distribution. The range of the values is divided into class intervals with equal widths, and the number of cases in each class interval is displayed. If a grouping variable is selected with unequal interval widths, the histogram will adjust the heights of the histogram bars to take account of unequal widths. The system gives you the flexibility of defining your intervals through cutpoints, this avoids creating histograms with zero counts on certain intervals.

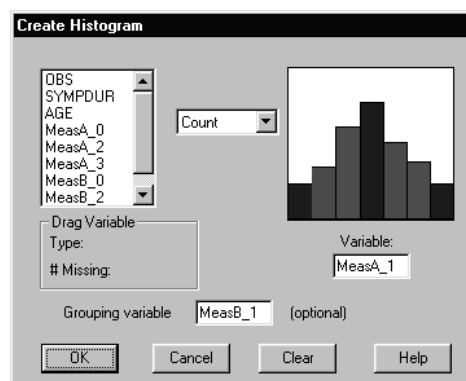
Frequencies, percentages or proportions may be displayed. A normal curve may be overlaid on the histogram for a visual test of normality.

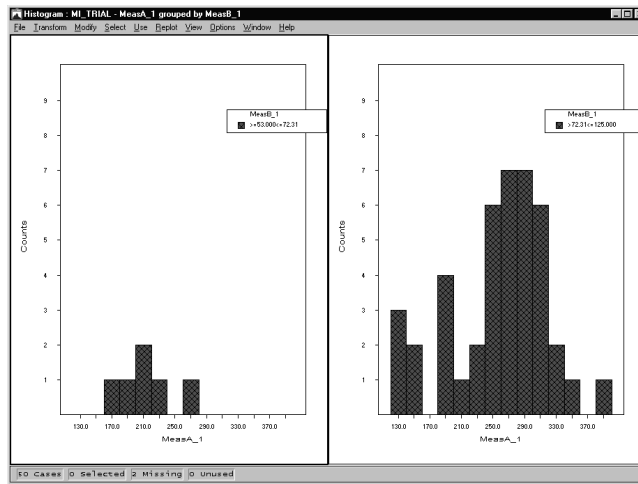
The histogram is also available with the square root of the frequencies plotted on the vertical axis. This special histogram is called a rootogram. The square root is an approximate variance stabilizing transformation for the binomial distribution. The standard deviation of the height is approximately proportional to the square root of the expected height. Here, the heights of the bars can no longer be directly interpreted as frequencies. See Velleman and Hoaglin (1981).

Creating a Histogram

Select Histogram from the Datasheet **Plot** menu, the system displays the Create Histogram window. The listboxes and datafields are described below:

Variables Listbox	Displays the list of used variables in the open datasheet.
Type	Displays the type of the variable currently selected.
Role	Displays the role of the variable currently selected.
Vertical Axis type	
Scrolled datafield	You can display the vertical axis type as: Count, Percent or Proportion.
Variable datafield	Drag and drop a variable from the Variables listbox to the Variables datafield.
Grouping Variable datafield	If you enter a variable and it is not grouped, the system will prompt you to group the variable.
OK button	The Y variable must be specified before the OK button becomes available.





Entering the required variables and pressing the **OK** button in the Create Histogram window displays an output window as shown above.

Boxplot

Boxplot is a popular method of displaying a univariate ordinal or continuous distribution. It was developed by Tukey and is useful both looking at a distribution from a single sample, and in comparing different distributions based on a number of samples simultaneously. It is a visual tool that utilizes the five-number summary statistics (minimum, first quartile, median, third quartile and maximum) in describing location, spread and shape of a distribution. Boxplots are recommended when comparing characteristics of different distributions.

The Boxplot displayed in the system follows definition 4 for quartiles with $k = 1.5$ (k is a factor for determining location of the fences). See Frigge *et al.* (1989) for definition. The upper and lower sides of the box enclose the middle half of the data and are the upper (Q3) and lower (Q1) quartiles, respectively. The length of the box is equivalent to the interquartile range ($IQR = Q3 - Q1$). The upper fence is at the maximum value or $(Q3 + 1.5(IQR))$, whichever is lower, and the lower fence is at the minimum value or $(Q1 - 1.5(IQR))$, whichever is higher. Values falling outside of the fences are called outside values and may possibly be outliers when the underlying distribution is normal.

The system includes an option to display a confidence interval for the median based on ordered statistics.

A $(1-\alpha)$ 100% confidence interval is defined by:

$$[x_{(i)}, x_{(j)}]$$

where: $x_{(i)}$ is the i th ordered value. The location of the confidence limits in the ordered data is

$$i = [(N+1)/2] + (\sqrt{N}/2)\Phi^{-1}(\alpha/2)$$

and: $j = N+1-i$

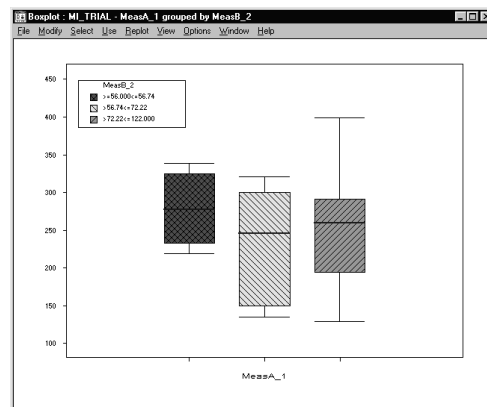
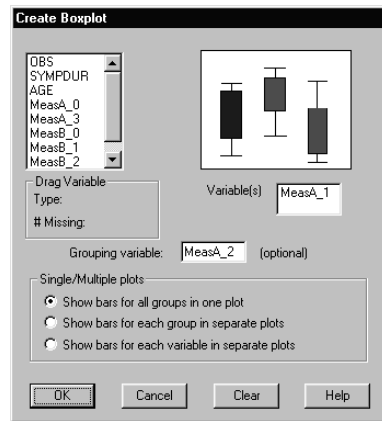
where: N is the sample size, and $\Phi^{-1}(\cdot)$ is the inverse cumulative normal distribution function.

The locations i and j may be non-integer-valued. In such cases, linear interpolation is used. This derivation is based on asymptotic normal approximation, and may be inappropriate for small sample sizes. Refer to “An Introduction to Statistical Analysis” - Dixon and Massey.

Creating a Boxplot

Select **Boxplot** from the Datasheet **Plot** menu, the system displays the Create Boxplot window. The listboxes and datafields are described below:

Variables Listbox	Displays the list of used variables in the open datasheet.
Type	Displays the type of the variable currently selected.
Role	Displays the role of the variable currently selected.
Variable(s) datafield	Drag and drop variables from the variables listbox to the Variables datafield.
Grouping Variable datafield	If you enter a variable and it is not grouped, the system will prompt you to group the variable. Once a grouping Variable is selected, other relevant options become available. If you choose a variable which is not grouped, the system prompts you to group the selected variable.
Show Bars for all groups in one plot checkbox	Lets you customize your plot and show bars for all groups in one plot denoted by different symbols. It is grayed out unless you have chosen a grouping variable.
Show bars for each group in different plots checkbox	Lets you customize your plot and show bars for each group in different plots It is grayed out unless you have chosen a grouping variable.
Show bars for each variable in different plots checkbox	Lets you customize your plot and show bars for each variable in different plots. It is grayed out unless you have chosen a grouping variable.
OK button	The Y variable must be specified before the OK button becomes available.



Entering the required variables and pressing the **OK** button in the Create Boxplot window displays an output window as shown above.

Bar Chart

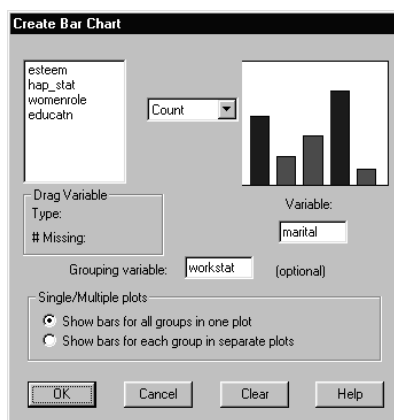
Bar chart is commonly used to display categorical variables. A bar represents each category, or level, and the height of the bars represents the frequency of cases under that category. The system allows you to specify the vertical scale either as counts, proportion, or percent. You may use a grouping variable.

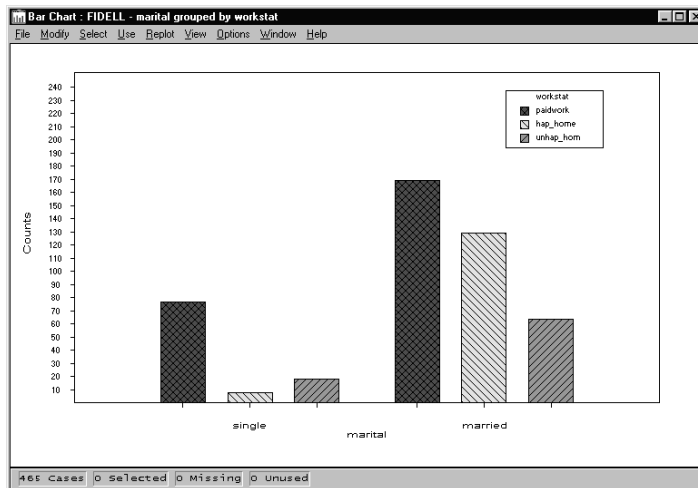
Creating a Barchart

Select **Bar Chart** from the Datasheet **Plot** menu. the system displays the Create Barchart window. The listboxes and datafields are described below:

- Variables Listbox** Displays the list of used variables in the open datasheet.
- Type** Displays the type of the variable currently selected.
- Role** Displays the role of the variable currently selected.

Variable datafield	Drag and drop variables from the Variables listbox to the Variables datafield.
Grouping Variable datafield	If you enter a variable and it is not grouped, the system will prompt you to group the variable. Once a grouping Variable is selected, other relevant options become available.
Show bars for all groups in one plot checkbox	Lets you customize your plot and show bars for all groups in one plot denoted by different symbols. It is grayed out unless you have chosen a grouping variable.
Show bars for each group in different plots checkbox	Lets you customize your plot and show bars for each group in different plots. It is grayed out unless you have chosen a grouping variable.
OK button	The X variable must be specified before the OK button becomes available.





Entering the required variables and pressing the **OK** button in the Create Bar Chart window displays an output window as shown above.

Means Comparison Chart

Means comparison plot is useful in presenting a visual comparison of means from different groups or of means of different variables (for example, in repeated measures), or a comparison of different distributions. When inference on means are of interest, an interval based on a multiple of the standard error of the mean may be used; 1 standard error is the default. Assuming normality, use 2 standard errors for 95% confidence intervals. When intervals do not overlap, then there is evidence that the means are statistically different. When distributions are of primary interest, then the standard deviation of the cases may be used in constructing an interval. The plots are also useful for assessing homogeneity of variances for different groups or variables.

There are two types of means comparison in the system. Means are represented by a point-by-point default with whiskers representing some measure of uncertainty. Alternatively, means may be represented by the height of bars or bins. The distance from the height of the bar to the top of the error bar measures Dispersion.

Create Means Comparison Chart

Select **Means Comparison Chart** from the Datasheet **Plot** menu. The system displays the Create Means Comparison Chart window. The listboxes and datafields are described below:

Variables Listbox Displays the list of used variables in the open datasheet.

Type Displays the type of the variable currently selected.

Role Displays the role of the variable currently selected.

Chart Type checkbox Lets you specify Bins or Points as the desired chart type.

Variable(s) datafield Drag and drop variables from the variables listbox to the Variable(s) datafield.

**Grouping Variable
datafield**

If you enter a variable and it is not grouped, the system will prompt you to group the variable.

**Show Bars for all
groups in one plot
checkboxbutton**

Lets you customize your plot and show bars for all groups in one plot denoted by different symbols. It is grayed out unless you have chosen a grouping variable.

**Show bars for each
group in different
plots checkboxbutton**

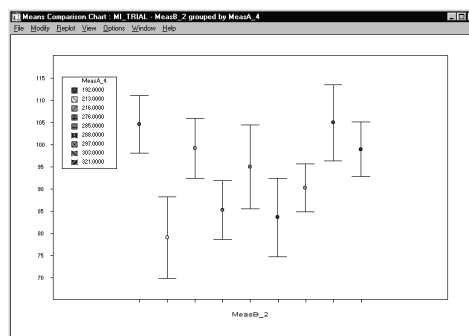
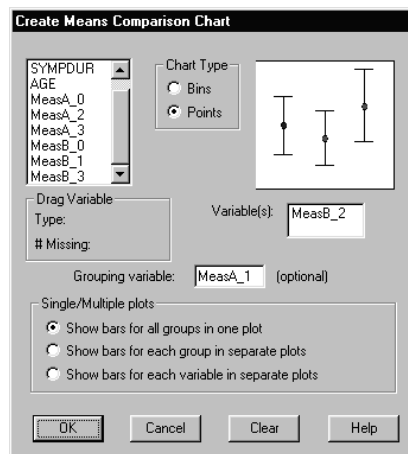
Lets you customize your plot and show bars for each group in different plots. It is grayed out unless you have chosen a grouping variable.

**Show bars for each
variable in different
plots checkboxbutton**

Lets you customize your plot and show bars for each variable in different plots. It is grayed out unless you have chosen a grouping variable.

OK button

The Y variable must be specified before the **OK** button becomes available.



Entering the required variables and pressing the **OK** button in the Create Means Comparison window displays an output window as shown above.

Normal Probability Plot

Normal probability plot is used to show graphically whether a variable is normally distributed or not. The numerical values of a variable are plotted on the horizontal axis, and transformed cumulative proportions (from the cumulative standard normal distribution) are plotted on the vertical axis. The transformation is such that if the data are normally distributed the points will lie on a straight line.

In the system, the data values are ordered before plotting and the expected normal value based on the ranks is plotted on the vertical axis. The expected normal value of the j th ordered value is:

$$\Phi^{-1}([3j - 1] / [3N + 1])$$

where: N is the sample size, and $\Phi^{-1}(\cdot)$ is the inverse cumulative normal distribution function.

If the extremes of the normal probability plot are curved downward, this may be an indication of a distribution that is skewed to the right. A distribution that is skewed to the left will have the extremes of the normal probability plot curved upward. An S-shaped normal probability plot may indicate that the distribution has heavier tails than a normal.

Reference lines may be included to further detect deviation from linearity. The normal reference line displays a line defined by:

$$y = -\frac{\bar{x}}{s} + \frac{1}{s}x$$

where: \bar{x} and s are the sample mean and standard deviation, respectively.

To assess symmetry, a robust reference line may be used. This line passes through the first and third quartile points to the data. If the line also intersects the median value and the lower half of the data is a mirror image of the upper half, then the empirical distribution is symmetric with respect to the median or mean.

An alternative display is the detrended normal probability plot. The linear trend is removed by subtracting the standardized score of the ordered value from its expected normal value. A relatively flat pattern around zero would suggest normality. Large differences at either or both ends indicate skewness or long tailedness.

Another option is the half-normal probability plot. This is used when assessing whether your data came from a standard half-normal distribution or not. The half-normal density results from a standard normal density being folded at 0 (the mean of the standard normal). Only positive values are allowed and thus, if your data contain negative values, absolute values are obtained before the data are sorted. The expected half-normal value of the j th ordered value is approximated by:

$$\Phi^{-1}([3N + 3j - 1] / [6N + 1])$$

where: N is the sample size, and $\Phi^{-1}(\cdot)$ is the inverse cumulative normal distribution function.

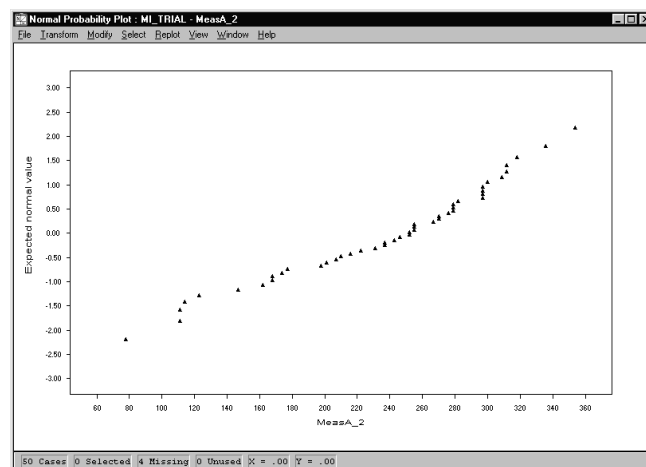
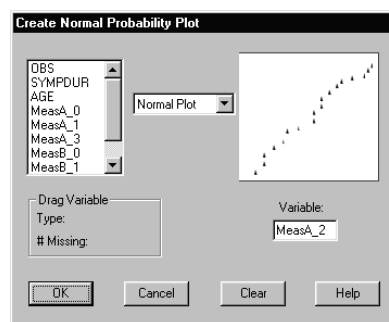
When half-normal plot is used as a test for normality, always standardize your variables such that the mean of the used variables is zero.

A linear pattern is indicative of normality. This plot option may be readily used in regression diagnostics where an assessment of normality of residuals is desired. The residuals need not be standardized, since their mean is already zero.

Create Normal Probability Plot

Select **Normal Probability Plot** from the Datasheet **Plot** menu. The system displays the Create Normal Probability Plot window. The listboxes and datafields are described below:

Variables Listbox	Displays the list of used variables in the open datasheet.
Type	Displays the variable type currently selected.
Role	Displays the role of the currently selected variable.
Scrolled datafield for Y axis	Select Normal Plot, Detrended Normal, or Half Normal Pot
Variable datafield	Drag and drop a variable from the variables listbox to the Variable datafield.
OK button	An X variable must be specified before the OK button becomes available.



Entering the required variables and pressing the **OK** button in the Create Normal Probability Plot window displays an output window as shown above.

8. Tutorial

SOLAS MAIN FEATURES WITH EXAMPLES

Introduction

This chapter provides a statistician's view of the system by discussing data screening. It includes discussions of system features and some examples to help you learn the system. The examples given cover the main features. We encourage you to follow these examples. Whether you try the examples or not, you will probably find the statistical discussions interesting.

Where to look for Information

You can access the online Help by pressing the Help button in any window. Also available is help for statistical analyses that includes both statistical discussions and definitions.

Screening and Changing Data in the system

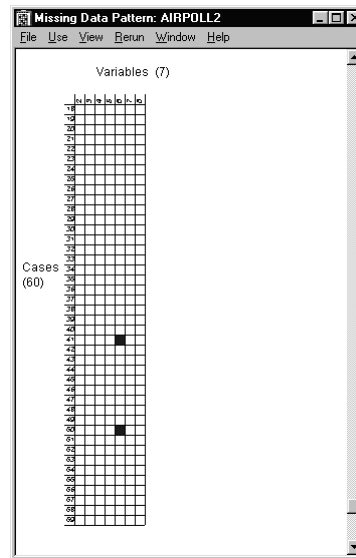
Before you start screening data, you should save one copy of your datasheet with a different file name, or use the **Copy Datasheet** option. If you do either, you can always go back to your original datasheet. While displaying the datasheet, click on the **File** menu **Save As** option. Replace the file name in the upper left-hand corner with a new file name. You now have two files of the data.

Missing Data Pattern

Many investigators first like to check the pattern of missing data in the datasheet. If data for a variable are missing in a high proportion of cases, then the use of that variable should be carefully considered. Analyses such as multiple regression use only cases that have complete data, so including a variable with numerous missing values can severely reduce the sample size.

To see the pattern of missing values:

1. In an open datasheet, click on the **View** menu **Missing Data Pattern** option. This option gives you a case by variable array of your data with the colored squares indicating where missing values occur.
2. Right-clicking and holding down the mouse button on any square displays a window containing the name of the variable and case for that square. It is easy to see which variables have numerous missing values.
3. Right-clicking and holding down the mouse button at the top of the column of an array brings up the name of the variable and its number of missing values. If the data are entered chronologically, you can see whether missing values are clustered in certain time periods.



See the SOLASTM manual for a discussion of the various imputation techniques available when you have the Imputation package. These techniques help you to deal with missing data.

Outliers

After dealing with missing data, many investigators next want to confirm that there are no outliers in the datasheet. Outliers are defined as cases that are inconsistent with the remainder of the data.

For nominal or ordinal data (discrete data), one obvious method is to examine the frequency distribution of each variable to find all undefined values. To do this, click on the **Analyze** menu **Descriptive Statistics** option. Choose the **Ordinal/Nominal** option.

You will get the total number and proportion of cases having a particular value. If a variable value occurs that should not occur, you should (if possible), trace the source of the data and check the validity of that value. In some cases it may be appropriate not to use that value. However, you should note that not all outliers are necessarily incorrect, and that looking at outliers can be very informative. Casually throwing outliers out of all analyses may bias the results.

Outliers in Small to Medium Sized Datasheets

For a small to medium sized datasheet, knowing that a particular variable has outliers allows you to identify the outliers easily. Then you can just click on the cell in the datasheet and delete the incorrect value. At that point, you can either leave the cell empty, or replace the value with a correct value.

Outliers in Continuous or Ordinal Data

For continuous or ordinal data, one quick method of finding variables with possible outliers is to obtain simple descriptive statistics for each variable by clicking on the **Analyze** menu

Descriptive Statistics option for **Continuous/Ordinal** data. If you already have an idea of the values that you can reasonably expect for maximum and minimum values, then scanning the list of maximum and minimum values will tell you if there is at least one case that is suspect for any given variable.

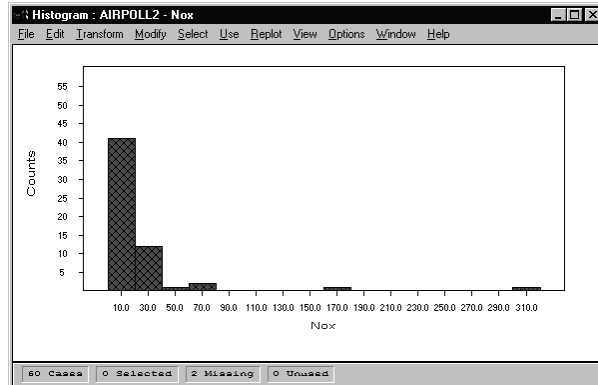
Looking at the maximum and minimum z -scores can be helpful in screening the data for outlying values and skewed or non-normal distributions. A potential rule of thumb is to take a closer look at any variable for which the absolute value of either the maximum or minimum z -score is greater than about 2.5. Of course, how extreme the z -score must be to signify an outlier depends on the sample size. In a very small sample an outlier may inflate both the mean and the standard deviation so much, that the outlier may not have a very extreme z -score. In a sample of several thousand cases, a maximum z -score of 3.5 would not be an unusual occurrence even for a normally distributed variable.

When outliers are suspected, it is also helpful to look at the distance of an observation not from the mean, but from the trimmed mean. A large skewness coefficient relative to its standard error may occur when an otherwise symmetric distribution has a large outlier, as well as when the sample distribution is in fact skewed.

Finding and Removing Outliers with the Histogram

To display a variable having possible outliers by using the Histogram, perform the following procedure:

1. Open the Airpoll2.mdd datasheet, select the **Plot** menu **Histogram** option.
2. Drag and drop the “Nox” variable from the displayed list of variables to the Variable datafield.
3. Click **OK** to display the histogram

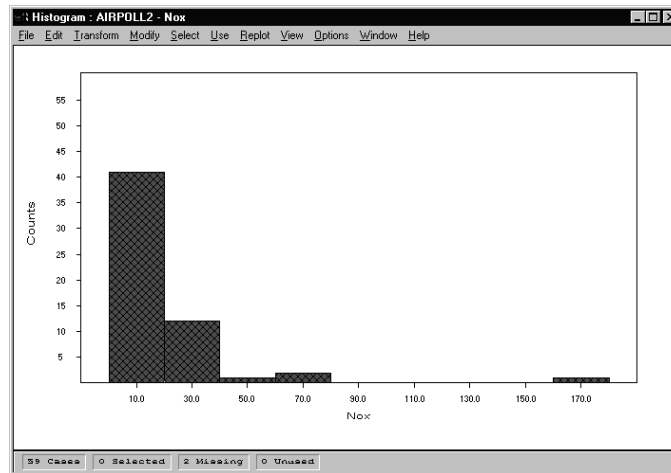


If one of the rectangles in the histogram is removed from the others, it may be the result of an outlier.

If you want to see the effects of removing the point(s) from the datasheet:

1. Click on the rectangle at the right-hand side of the histogram. Heavy lines appear surrounding the rectangle, and the status bar displays the number of cases selected.

NOTE: In the case of the Nox variable however, we also see marked evidence of skewness. When the variable is transformed by taking logs, the histogram of $\log(\text{Nox})$ no longer shows any marked evidence of lack of symmetry. The two apparent outliers are seen to be quite consistent with the rest of the distribution. It will often be helpful in assessing whether extreme values are outliers or not to investigate the effect of transforming the variable first.



2. Click on the **Use** menu **Do Not Use Selected Cases (Global)** option. This action causes the case(s) to be grayed out in the datasheet for that variable, and not available for use. The unused bar disappears from the histogram. A similar procedure works for scatterplots.

For continuous data, if you are concerned that you may have numerous outliers that you may not be able to remove, consider using the robust procedures available in the system.

How to Find Outliers — Example

1. Go to Start-up window and choose the **File** menu **Open** option.
2. Select Airpoll2.mdd and click the **OK** button. We will use this datasheet to look for missing values and outliers.

Airpoll2.mdd should now be displayed in the Datasheet window. Airpoll2.mdd has seven variables including the variable “Name” that contains the names of cities. The datasheet also includes six other variables that we will later try to use to predict mortality.

The variable definitions are:

- Annual rainfall in inches , median school year completed, population per square mile in 1960
- Percent non-white, relative pollution potential of Nox and SO₂, and total age-adjusted mortality or deaths per 100,000 people.

	Name	Rain	Pop_den	Nonwhite	Nox	So2	Mortality
1	akronOH	36	3243	8.8	15	59	921.9
2	albanyNY	35	4281	3.5	10	39	997.9
3	allenPA	44	4260	0.8	6	33	962.4
4	atlantGA	47	3125	27.1	8	24	982.3
5	baltimMD	43	6441	24.4	38	206	1071.0
6	birmnghAL	53	3325	38.5	32	72	1030.0
7	bostonMA	43	4679	3.5	32	62	934.7
8	bridgeCT	45	2140	5.3	4	4	899.5
9	buffaloNY	36	6582	8.1	12	37	1002.0
10	cantonOH	36	4213	6.7	7	20	912.3
11	chataTN	52	2302	22.2	8	27	1018.0
12	chicagoIL	33	6122	16.3	63	278	1025.0
13	cincinnatiOH	40	4101	13.0	26	146	970.5
14	clevelandOH	35	3042	14.7	21	64	986.0
15	columbusOH	37	4259	13.1	9	15	958.8
16	dallasTX	35	1441	14.8	1	1	860.1
17	daytonOH	36	4029	12.4	4	16	936.2
18	denverCO	15	4824	4.7	8	28	871.8
19	detroitMI	31	4834	15.8	35	124	959.2

The data set Airpoll.mdd is described in Appendix A: Data Sets.

NOTE: To scroll, click on the arrows in the horizontal scroll bar at the bottom of the window and in the vertical scroll bar to the right. Another way to scroll sideways is to click and hold down the mouse button while the cursor is on the square button in the horizontal scroll bar, then drag it to the left or right.

- To view the missing data locations, click on the **View** menu, then **Missing Pattern**.

Missing Value Pattern: AIRPOLL2

Variables (8)

Cases (60)

Case: 41
Variable: Nox
Status: Missing

Pairwise Missingness Report

	Name	Rain	Education
Pop_den	0	0	0
Nonwhite	0	0	0
Nox	2	2	2
So2	0	0	0
Mortality	0	0	0

Pairwise Presence Report

	Name	Rain	Education
Name	60		
Rain	60	60	
Education	60	60	60
Pop_den	60	60	60

- Right-click and hold on the upper filled-in square to display a message that “Case 41 for variable Nox is missing”. Similarly, if you click and hold on the second filled-in square, you can see that case 50 for variable Nox is also missing

NOTE: Selecting the **View** menu, then **Pairwise Missingness Report** shows that the two missing cases for the Nox variable will be excluded for every variable in the datasheet in a pairwise analysis.

- To bring up the total number of missing cases for the Nox variable, click and hold at the top of the Nox variable column. There are no other missing values in the datasheet, so we know that missing values are not a major problem in this datasheet.

Let's try to find some outliers.

- Close the Missing Data window.
- Click on the **Analyze** menu again in the Datasheet window.
- Choose **Descriptive Statistics** and the **Continuous/Ordinal** option.

Descriptive Statistics (Cont/Int/Ord) : AIRPOLL2

File Edit Options Window Help

Arial 10 B I U

DESCRIPTIVE STATISTICS

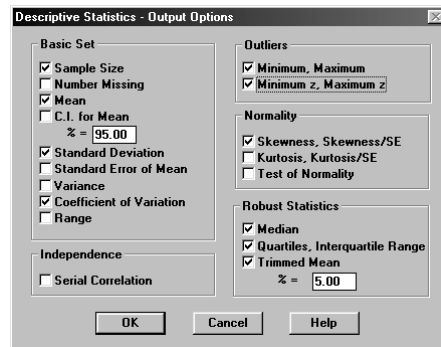
	N	Mean	StdDev	Min
Rain	60	39.0333	15.3169	10.0000
Education	60	10.9733	0.8452	9.0000
Pop_den	60	3818.7700	1532.1515	315.2000
Nonwhite	60	11.8700	8.9211	0.8000
Nox	58	22.9482	47.0920	1.0000
So2	60	53.7666	63.3904	1.0000
Mortality	60	940.3816	62.2124	790.7000

	Max	Median
Rain	128.0000	38.5000
Education	12.3000	11.0500
Pop_den	9699.0000	3567.0000
Nonwhite	38.5000	10.4000
Nox	319.0000	9.0000
So2	278.0000	30.0000
Mortality	1113.0000	943.7000

Number of cases used : 60

NUM Col: 0 Line: 91 Page: 1

- The default output does not include Skewness, Coefficient of Variation, Minimum and Maximum z -values, or Trimmed Mean.
- To add those statistics to the output screen, click on **Options**, and choose the **Output** option. The Descriptive Statistics Output Options window is displayed.
- Check the Skewness, Coefficient of Variation, Minimum and Maximum z -values, and the Trimmed Mean option, then click the **OK** button.



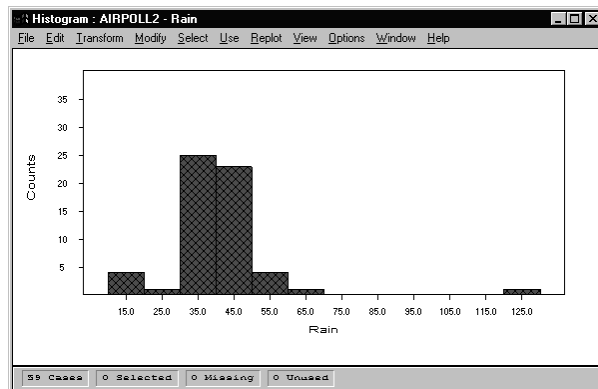
Note that the Minimum and Maximum z -values are not symmetric, and there are some large values. Large maximum z -scores may be expected if the data are skewed to the right. Rain, Pop_den, Non-white, Nox, SO₂, and Mortality all show large maximum z -values. Data are seldom skewed to the left, so very small minimum z -scores are more apt to indicate a small outlier. Education, Pop_den, and Mortality all show somewhat small minimum z -values. We probably have some non-normal data and/or outliers.

We can see that there are some large values for skewness, especially for Rain, Nox and SO₂. This is a datasheet that needs screening.

Let us look at another simple measure, the difference between the Trimmed Mean and the Maximum and Minimum values. The maximum of 128 looks large for Rain; 315 looks small and 9699 looks large for Pop_den; 319 looks large for Nox; and 278 looks large for SO₂.

We can look further at the Rain variable.

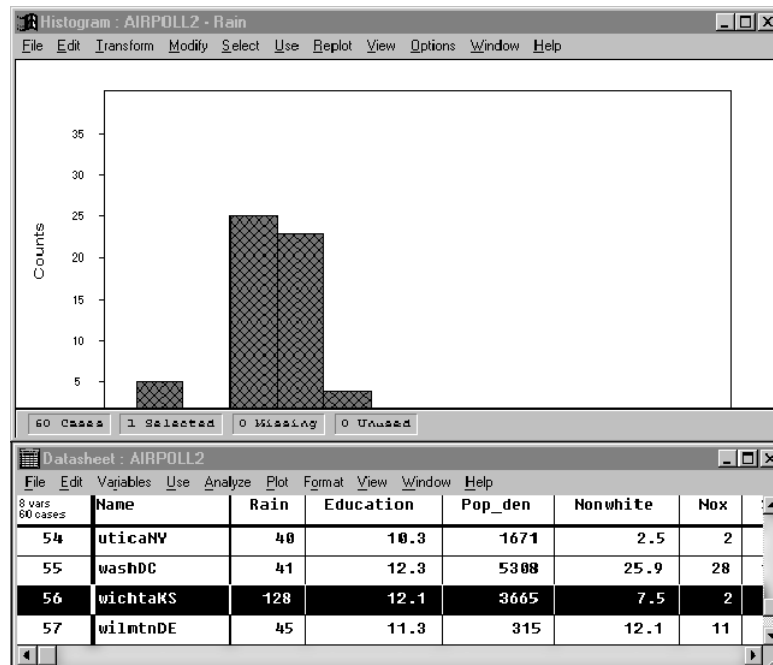
1. Close the descriptive statistics window by going to the **File** menu and clicking on **Close**.
2. From the **Plot** menu select the **Histogram** option.
3. Drag and drop the variable Rain to the Variable datafield.
4. Click **OK**.



When you look at the histogram, the bar centered at 128.0 is clearly far from the rest of the values. We could eliminate that value now, or we can look up the associated city, and then decide. But it is a candidate outlier.

To find the city quickly:

1. Click on the outlying bar at 128.0.
2. Select the **Window Tile** option. You can view the histogram and scroll through the datasheet to find the city highlighted by your selection of the 128.0 box.



Now let us try Pop_den, repeating the same steps. This situation is less clear, but the upper value is somewhat suspect. We can look at the maximum value and the minimum value. Knowledge of the city may help here.

A similar trial of Nox shows a distribution skewed to the right with two large values. Finally, an SO₂ histogram also shows two large values. Select those cases in the histogram, and then they are highlighted in the datasheet so that you can find them easily. After highlighting, go back to the datasheet.

Scrolling down in the datasheet, you find that heavy rain occurs in Wichita, Kansas. This does not seem correct, so this value should be removed or corrected. The population density of Wilmington, Delaware appears to be unreasonably small, so it should be corrected or removed. The large value for York, Pennsylvania is more of a puzzle. Finally, the large value for Nox occurs in Los Angeles, so maybe that is a real but very large value. San Francisco is also large. For SO₂, one of the large values is Pittsburgh and another is Chicago. Those values could be real or not.

In this example, we have two clear outliers and several suspicious values. This is typical in screening for outliers; some cases are clear, and some are indecisive. Let us remove the clear ones (heavy Rain and small Pop_den) and leave the others for the time being.

One straightforward way of removing outliers is to delete those observations from the datasheet. In each case, use the mouse to click on the cell to be deleted, then use the <Delete> key. Save this adjusted file as the new file with a different name.

NOTE: Remember that your report should inform readers that values have been deleted. Also, you should provide references for the statistical methods used in deciding the outlying values to be deleted.

Logical Inconsistencies

Logical inconsistencies in the datasheet may also be a sign of incorrect data. For example, pregnant males should not occur. You can use the two-way frequency table to check for known inconsistencies.

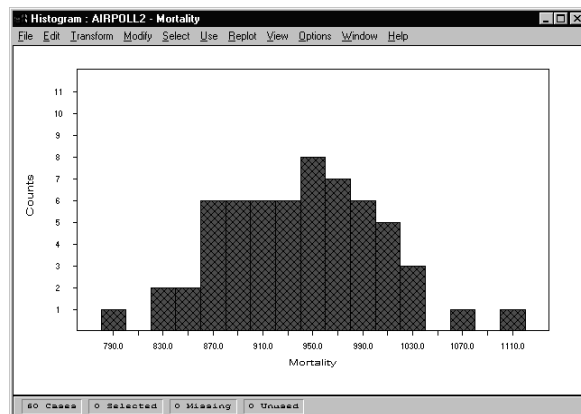
Normality

In addition to screening the data for missing values and outliers, we also want to check whether our continuous data are at least approximately normally distributed. When we know that, we will know if the data meet the assumptions for the various tests and confidence limits available to us. If the data are not normally distributed, we may be able to transform the data to a near-normal distribution. The system has numerous tools for checking normality. Normal distributions are symmetric and have a bell-shaped curve.

Checking for Normality with Histogram

To check the data distribution using a Histogram:

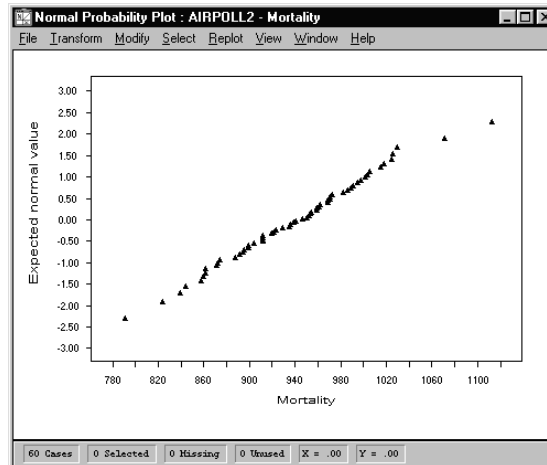
1. Click on the **Plot** menu and select **Histogram**.
2. Drag and drop a variable from the variables list to the Variable datafield, then click the **OK** button to display the plot.



Viewing the histogram makes it easy to see if the variable is symmetric. It is sometimes difficult to judge symmetry and normality when the sample is small, and/or outliers are still present.

Checking for Normality with Normal Probability Plot

When checking for normality, the Normal Probability plot is another widely-used method. If you click on the **Plot** menu in the Datasheet window, you can select the **Normal Probability Plot** option. Then, drag and drop a variable from the Variables list to the variable datafield. Press the **OK** button to display a normal probability plot.



Normally Distributed Data

If the plot looks like a straight line except for the extreme points, then for all practical purposes you can assume the data to be normally distributed.

Skewed Data

If the plot looks like a curve with the ends pointed downward, then the data are skewed toward the right. If the ends of the curve point upwards, then the data are likely skewed to the left. Additional shapes are shown in Afifi and Clark (1990), pages 60-61.

Statistics Used in Checking for Normality

Several statistics are also useful in detecting lack of normality.

1. Starting from the datasheet again, click on the **Analyze** menu.
2. Choose **Descriptive Statistics** and **Continuous/Ordinal** to get the Descriptive Statistics output.
3. Go to the **Options** menu and choose the **Output** option.
4. Check the Skewness, Kurtosis, and Test of Normality options, then click the **OK** button.

In the updated output window, you may have to scroll to the right to find the output, because it appears to the right of the usual statistics.

Skewness and kurtosis are measures of asymmetry and long-tailedness of the distribution of a variable. For a normal distribution, the expected value is zero for both of these statistics. Both measures are very sensitive to outliers, and are difficult to interpret if outliers are present. They also may be highly variable for small sample sizes.

The ratio of skewness to its standard error can be roughly read as a standardized z -value.

Positive values greater than 2 or 2.5 are unusual, and may indicate a distribution skewed to the right. Similarly, small negative values indicate a distribution skewed to the left. Large values (greater than 2 or 2.5) of kurtosis divided by its standard error indicate a long-tailed distribution.

The output also includes the W test developed by Shapiro and Wilk. The null hypothesis being tested is that the observations are from a normal sample. This test is especially useful for small samples where some of the other methods are not very good. The W statistic is positive with a maximum value of 1. A small W indicates departure from normality. The actual significance level is given in the column to the right of the statistic.

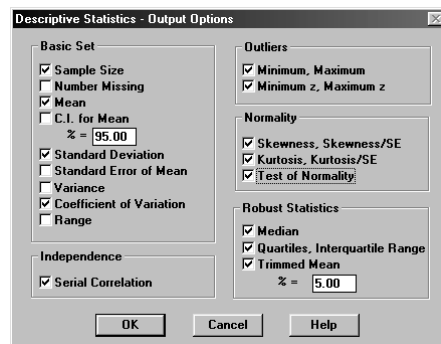
Checking for Independence

Finally, you may want to determine whether the cases in the sample are independent. In many studies where observations are collected on separate individuals or objects, this is not a problem. However, lack of independence may occur if equipment is drifting out of calibration, or if the data are collected over time, and time has an effect on the outcome.

NOTE: Multiple measures on the same subject cannot be considered independent.

Statistics Used in Checking for Independence

1. In the Descriptive Statistics window, go the **Options** menu and choose the **Output** option.
2. Check the Serial Correlation option.



3. Click on **OK**. This will put you back in the Descriptive Statistics output window. You may have to scroll to the right to find the output, because it appears to the right of the usual statistics.

Descriptive Statistics (Cont/Int/Ord) : AIRPOLL2

File Edit Options Window Help

Arial 10 B I U

DESCRIPTIVE STATISTICS

	N	Mean	StdDev	Min
Rain	60	39.0333	15.3169	10.0000
Education	60	10.9733	0.8452	9.0000
Pop_den	60	3818.7700	1532.1515	315.2000
Nonwhite	60	11.8700	8.9211	0.8000
Nox	58	22.9482	47.0920	1.0000
So2	60	53.7666	63.3904	1.0000
Mortality	60	940.3816	62.2124	790.7000

	Max	SerCorr	p-value	Median
Rain	128.0000	0.2193	0.0445	38.5000
Education	12.3000	0.0379	0.3843	11.0500
Pop_den	9699.0000	-0.0524	0.6577	3567.0000
Nonwhite	38.5000	0.0152	0.4529	10.4000
Nox	319.0000	0.1051	0.2116	9.0000
So2	278.0000	0.2300	0.0373	30.0000
Mortality	1113.0000	0.1225	0.1712	943.7000

Number of cases used : 60

NUM Col: 0 Line: 108 Page: 1

The serial correlation is the product moment correlation between cases when each case is lagged one case behind the other. Thus, the value obtained will depend on the order of the cases in the data file. The significance level is the result of a test of the null hypothesis that the serial correlation is zero.

Serial correlation has little meaning for this datasheet, since the data for all of the cities in one region could have been entered into the datasheet at the same time. None of the example data sets represent time-series data to provide an example in which we can easily interpret the meaning of a significant serial correlation. All of the serial correlations appear to be small and non-significant, except for Rain and SO₂.

Let's check for normality and independence in one of the variables in the adjusted air pollution datasheet, Airpoll2.mdd.

Remember that when we obtained the histogram for Nox that it was skewed to the right. The data do not appear to be normally distributed.

Descriptive Statistics (Cont/Int/Ord) : AIRPOLL2

File Edit Options Window Help

Arial 10 B I U

DESCRIPTIVE STATISTICS

	Skewness	SkewSE	Kurtosis	KurtSE
Rain	2.9232	9.2442	16.7998	26.5628
Education	-0.2136	-0.6761	-0.8616	-1.3623
Pop_den	1.0049	3.1778	2.5165	3.9790
Nonwhite	1.0752	3.4000	0.6370	1.0073
Nox	4.8196	14.9654	25.6558	39.8835
So2	1.8172	5.7467	2.9603	4.6807
Mortality	0.0913	0.2889	-0.0553	-0.0874

	W Stat	p-value
Rain	0.7377	6.5647E-14
Education	0.9488	0.0270
Pop_den	0.9472	0.0221
Nonwhite	0.8914	1.2674E-05
Nox	0.4389	1.2011E-25
So2	0.7631	1.2307E-12
Mortality	0.9927	0.9946

Number of cases used : 60

NUM Col: 0 Line: 91 Page: 1

Scroll the output window to the extreme right. Note that Nox has a highly significant ratio for both skewness and kurtosis.

The Shapiro and Wilk's W statistic significance level is near zero, a strong indication of non-normal data. In fact, Rain, Pop_den, Non-white, Nox and SO₂ all appear to be skewed to the right and clearly not normal.

Education does not have a significant skewness or kurtosis, but the W statistic is statistically significant at the 5 percent level. Rain is not normal according to the W test. If one is very concerned about having normal data, this datasheet will need some work.

Making Changes to Facilitate Using Data in an Analysis

Making changes to a variable for one particular case is very simple.

1. In the datasheet window, click on a cell with the mouse, and delete the contents.
2. At that point, you can type in a new value or leave the cell empty.
3. Click on the next cell to be changed.
4. After your last change in the datasheet, press <Enter>.

Transformations

Suppose we have some variables that are not normally distributed, and you want to transform these variables to obtain a more normal distribution. You can do this from the Datasheet window:

1. To use one of the given transformations, click on the name of the variable (at the top of the column) to be transformed in the datasheet.
3. From the datasheet, click on the **Variables → Transform** menu. The Transform menu displays a list of widely-used transformations, plus the **User-defined** option.
4. Click on the desired transformation. This will result in the transformed variable being placed to the right of the existing variables in the datasheet.

	Education	Pop_den	Nonwhite
1	11.4	3243	8.8
2			3.5
3			0.8
4	atlantGA	47	27.1
5	baltimMD	43	24.4
6	birmnhAL	53	38.5
7	bostonMA	43	3.5
8	bridgeCT	45	5.3
9	bufaloNY	10.5	6582
10	cantonOH	10.7	4213
11	chatagTN	9.6	2302
12	chicagIL	10.9	6122
13	cinnclOH	10.2	4101
14	cleveOH	11.1	3042

NOTE 1: If you select (highlight) a variable column in your datasheet, and then apply a simple transform from the displayed **Variables → Transform** menu, the system inserts the new variable to the right of the selected variable in your datasheet.

NOTE 2: If you select (highlight) a vertical line in any column in your datasheet, and then apply a simple transform from the displayed **Variables → Transform** menu, the new variable is inserted at that point.

Otherwise, the transformed variable is displayed in the column to the right of the last variable entered in your datasheet.

Custom Transformations

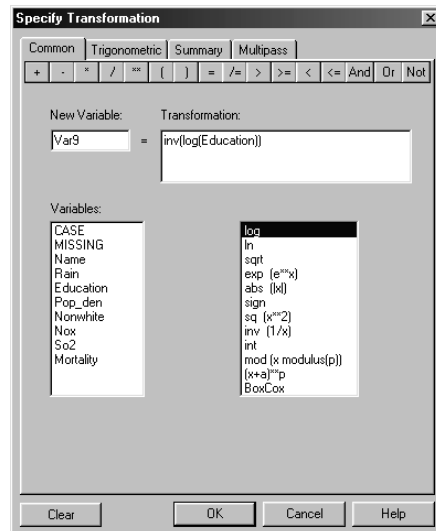
1. To make a custom transformation, do NOT select a variable before choosing the **Transform** option.
2. From the datasheet, click on the **Variables → Transform** menu, then click on **User-defined** to display the Transform window.

The Transform window has four tabs which when pressed display the following views:

- Common
- Trigonometric
- Summary
- Multipass

NOTE: Transformations require matching parentheses in the expression.

3. Suppose you want to take the inv log of a displayed variable. Click on **inv**, click on **log**, then click on the variable in the Variables list.



4. You will see that you have `inv(log(<selected variable>` in the transformation box. This is an incomplete formula because the left facing and right facing parentheses must be equal in number. You must enter the missing closing parentheses.

Since the cursor is at the end of the expression, just type `)` to produce the correct formula, **`inv(log(<selected variable>))`**.

5. When your transformation is satisfactory, press **OK** to close the window and create the transformation in the datasheet. If the **OK** button does not work, the program does not recognize your formula as being correct, and you should change the expression.

NOTE: While trying transformations, finding the optimum transformation for the sample does not automatically give you the optimum transformation in the population, which is what you really want. In general, the effect of non-normality will be less if the ratio of the standard deviation to the mean is small. If, for example, this ratio is 1.18 in the original variable, while the ratio is only 0.24 in the transformed variable, the transformed value indicates that the effect of non-normality will probably not be too serious.

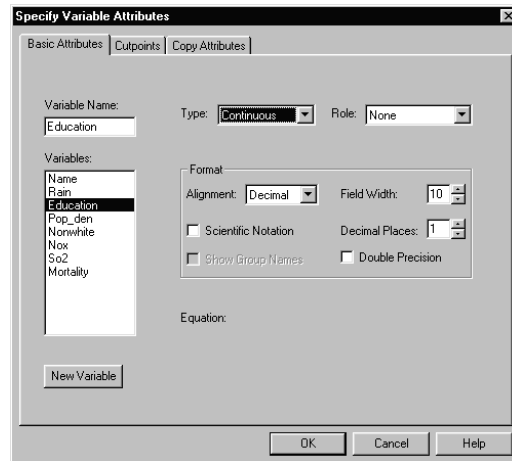
It is easy to do data screening and editing in the system. After a very short while, you can save a lot of time and have a carefully edited datasheet.

For more information about transformations, refer to Chapter 1 Data Management – “Transforming Variables” in the System Manual.

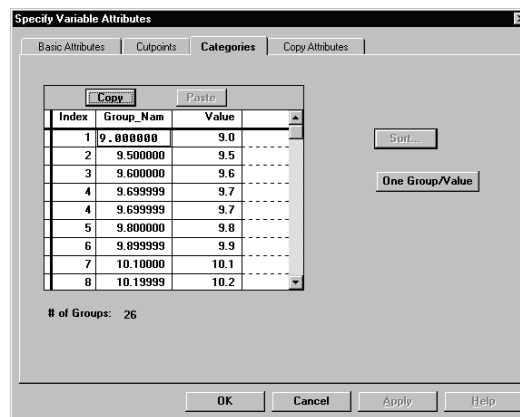
Combining Categories

Suppose you have a nominal variable with three categories. But the first category contains only one case, so you want to combine the first category with the second one. To combine categories:

1. Double click on the variable name at the top of the datasheet. The Variable Attributes window appears. For the Education variable, we have changed the Variable Type from Continuous to Nominal.

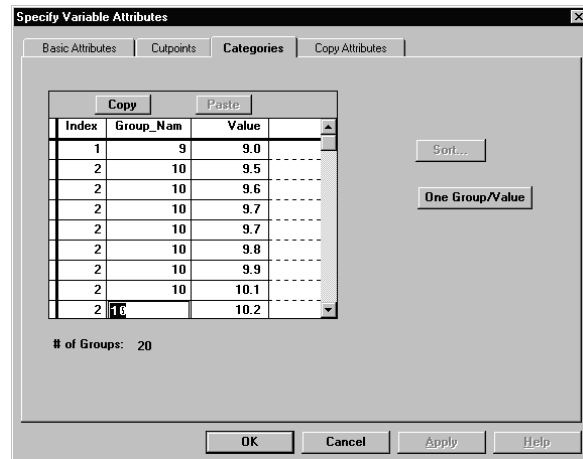


2. Click on the **Group** menu and choose the **Modify Categories** option. If you have only one nominal variable, the Modify Categories window will be filled with the information for that nominal value.



Remember that Education was really a continuous variable with values extending to one decimal place. When we changed it to a nominal variable, the values became integer values, as shown in the Values column of the Modify Categories window. However, the default name for each indexed case stems from the exact value of that case.

3. To change a Group_Name in the Modify Categories window, click on the group name. That name will be highlighted. Type in the new group name and press **<Enter>**. Continue changing group names through all the categories.



Notice that the group index is now the same for the three categories, although the numbers in the Value column have not changed.

4. Click on **OK** to go back to the Variable Attributes window.
5. Click on **OK** to return to the datasheet.
6. To view the new grouping in the datasheet, click on the nominal variable to highlight it. Select the **Variables** menu **Variable Attributes** option.
7. Check the Show Group Names box.
8. Click on **OK** to go back to the datasheet.

Statistical Data Analysis

Performing the analyses is a straightforward and simple process. You simply display the **Analyze** menu in the datasheet, and then select from the menu. You should try each of the available options.

When you want to set your preferences for output, you can do so from the SOLAS 3.0 Main window **View** menu **System Preferences** option, or from an Output window.

Descriptive Statistics

The **Descriptive Statistics** option provides statistics for two types of data: Continuous/Ordinal and Ordinal/Nominal (discrete) data. It also provides the serial correlation.

Descriptive Statistics for Continuous/Ordinal Data

To access the output, click on the **Analyze** menu in the datasheet. Choose the **Descriptive Statistics, Continuous/Ordinal option**. The Descriptive Statistics Output Options window is displayed.

To obtain the exact statistics that you want:

1. Click on the **Options** menu **Output** option.
2. In the Descriptive Statistics - Output Options window, click in the check boxes to make your choices. The options are organized by likely application so that you can find them easily.
3. If you want to set the confidence intervals to a different confidence level, check the **CI for Mean** box, and change the value in the datafield. You can do the same for the Trimmed Mean option.
4. Click on **OK**.

Subgroups in Continuous/Ordinal Data

You can also obtain statistics for subgroups of the datasheet if you have a grouping variable.

1. Choose the **Options** menu again.
2. Click on the **Grouping** option. This option will work directly if the grouping variable(s) is a nominal/ordinal variable. You can get statistics for a single grouping variable, or for two grouping variables in turn, or cross-classified.

Categorizing a Continuous Variable

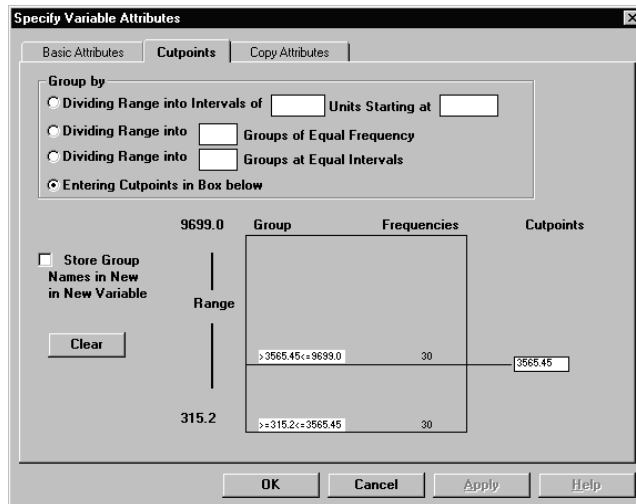
If you choose to categorize continuous data, you will use the graphical Set Cutpoints window. In this window you can see the range of your measurements for the chosen variable. You can easily add cutpoints one at a time, and you can drag any cutpoint to a different position.

To specify cutpoints for a grouping variable:

1. Click on a variable name in the datasheet. Choose the **Variables** menu **Group** option, then select **Set Cutpoints** to display the Set Cutpoints window.
2. The window provides four options. Most of the time you will find that choosing one of the first three options will work well for you. Click on your choice, then enter the required values in the respective datafield(s).

If you prefer, you can enter cutpoints directly in the Cutpoint graphics area.

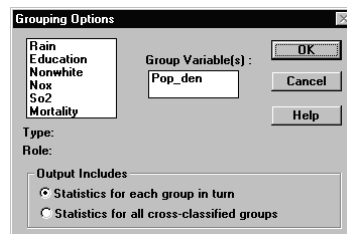
1. Click on the button to choose the **Entering Cutpoints in Box below** option.
2. Place the cursor in the Cutpoints area and click. A line is displayed.



3. The first cutpoint line separates the sample into two groups. For each group, the box displays the range and the frequency of cases in the group. Since your first attempt at a cutpoint is unlikely to be ideal, you can drag the line up or down. You can also click each time that you want to put in an additional line.
4. The graphical cutpoint box in the system gives you complete freedom to choose the exact cutpoints that you want. The program also lets you see the results instantaneously.
5. If you do not like your choice, click on **Clear** and try again.
6. When you are finished, click **OK**.

To view the descriptive statistics output with your new grouping:

1. Choose the **Analyze** menu **Descriptive Statistics** options for **Continuous/Ordinal**. The Descriptive Statistics window is displayed without grouping.
2. To add grouping, select the **Options** menu, choose the **Grouping** option, and select your new grouping variable.



3. Press **OK** to get the descriptive statistics output with grouping.

Descriptive Statistics (Cont/Int/Ord) : AIRPOLL2

File Edit Options Window Help

Anal 10 B I U

DESCRIPTIVE STATISTICS

Rain grouped by Pop_den

	Skewness	Skew/SE	Kurtosis	Kurt/SE
Rain	2.9232	9.2442	16.7998	26.5628
>=315.2 <=3513.03	-0.9705	-2.1701	0.6209	0.9178
>3513.03 <=9699.0	3.0761	6.8764	12.4211	13.8872

	W Stat	p-value
Rain	0.7377	6.5647E-14
>=315.2 <=3513.03	0.9173	0.0257
>3513.03 <=9699.0	0.6326	1.5263E-08

Education grouped by Pop_den

	Skewness	Skew/SE	Kurtosis	Kurt/SE
Education	-0.2138	-0.6761	-0.8616	-1.3623
>=315.2 <=3513.03	-0.1246	-0.2787	-1.0154	-1.1352
>3513.03 <=9699.0	-0.2690	-0.6016	-0.9769	-1.0922

NUM Col: 0 Line: 296 Page: 1

Descriptive Statistics for Ordinal/Nominal Data

For data of type ordinal/nominal (discrete or categorical):

1. Select the **Analyze** menu **Descriptive Statistics** → **Ordinal/Nominal** option. The frequencies and proportions are displayed for each nominal or ordinal variable.

You might want to use this output to find sub-categories containing frequencies that are insufficient to warrant further analysis.

You can find further information on using Descriptive Statistics in the Descriptive Statistics chapter of this manual. Definitions of the statistics appear throughout the manual.

Descriptive Statistics (Ord/Nom) : AIRPOLL2

File Edit Window Help

Anal 10 B I U

DESCRIPTIVE STATISTICS

Education: 60

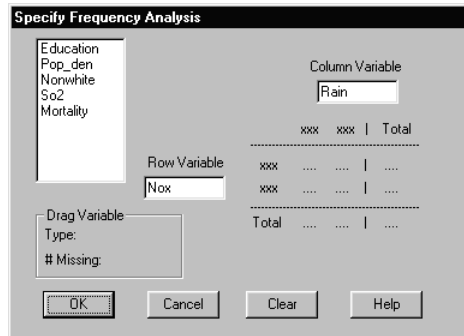
	N	Proportion
9.000000	1	0.0166
9.500000	1	0.0166
9.600000	0	0.0000
9.699999	3	0.0500
9.800000	2	0.0333
9.899999	1	0.0166
10.100000	1	0.0166
10.199999	2	0.0333
10.300000	2	0.0333
10.399999	1	0.0166
10.500000	4	0.0666
10.600000	0	0.0000
10.699999	2	0.0333
10.800000	4	0.0666
10.899999	2	0.0333

NUM Col: 0 Line: 108 Page: 1

Frequency Table Analysis

We discuss Two-way Frequency Analysis using the frequency table when you are starting from the datasheet.

1. To analyze a frequency table, start by choosing the **Analyze** menu **Tables** option. The Specify Frequency Analysis window is displayed. The window contains a two-frequency table with datafields for row and column variables.



2. Drag and drop variables to the Column Variable datafield, and the Row Variable datafield.
3. If you choose continuous variable(s), the system prompts you to group that variable(s). When you have grouped and placed both variables, click on **OK**. A window is displayed with the two-way table and the default statistics.

Frequency Output: AIRPOLL2 - Nox and Rain

File Edit View Options Format Window Help

Arial 10 B I U

TESTS

	Value	df	p-value
Pearson's Chi-square	5.9380	1	0.0148

TABLES

Table of Counts:

	>=10.00 <=33.73	>33.73 <=128.00	Total
>=1.00 <=79.17	13	43	56
>79.17 <=319.00	2	0	2
Total	15	43	58

NUM Col: 0 Line: 58 Page: 1

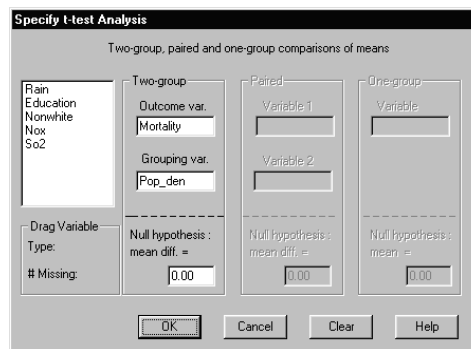
4. To change the output options, choose the **Output** option from the **Options** menu.
5. In the Output Options window, you can specify the types of Table, Measures, and Tests.

- Click on **OK**. Options that are not appropriate for the number of rows or columns in your table are grayed out.
- You can select **Options → Output** from the Output window, check Min. Expected Value in the Measures options, and press the **OK** button. You can then compare this new output value with the Chi-square value in the table. A computed value that is too small may indicate that the computed probability may not be accurate.

t- and Non-parametric Tests

You can use the *t*-test option for single sample, or two samples (groups) tests. The two sample analyses use either independent samples, or paired (matched) samples. To perform a *t*-test:

- In the Datasheet window, select the **Analyze menu → *t*- and Nonparametric Tests** option. The Specify *t*-test Analysis window is displayed.



- The Specify *t*-test Analysis window displays the datasheet variables and the choice of three *t*-tests: Two-group (independent), Paired, and One-group (single sample). Choose one of the three by dragging and dropping variable(s), from the list of variables, into the datafield(s) for the required *t*-test.
- After you drop a variable into one datafield, the datafields under the unused *t*-tests are grayed out. If you change your mind about a variable, you can drag that variable from the *t*-test datafield back to the list of variables.

You can select two types of variables:

- Outcome (dependent variables) and Grouping variables. For a Two-group test, you need one Outcome variable, and a Grouping variable that divides the sample into two groups.
- For a Paired *t*-test, you need two related Outcome variables, such as weight before dieting and weight after dieting.
- For the One-group test, you need only one Outcome variable.

If the grouping variable has more than two groups, or is ungrouped, you are prompted to group/re-group that variable. When you choose the **Group** option, the Set Cutpoints window is displayed.

- After grouping/regrouping the variable(s), click on **OK**, and you return to the *t*-test window. The grouping variable is displayed in the previously selected datafield

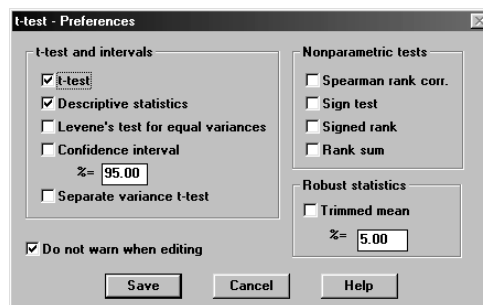
5. When you are satisfied with your selections, press the **OK** button, and the system displays the results for the *t*-test. If you had selected a Nonparametric test output, it is displayed at the end of the *t*-test results.

The screenshot shows a window titled "Two-group t and Non-param Tests : AIRPOLL2 - Mortality and Pop_den". It contains two tables: "DESCRIPTIVE STATISTICS" and "Test Statistics".

	N	Mean	StdDev	StdErr
Mortality	60	940.3816	62.2124	8.0315
>=315.2 <=3513.03	30	934.2400	67.5727	12.3370
>3513.03 <=9699.0	30	946.5233	56.8340	10.3764
Diff. of Means		-12.2833	N/A	16.1205

	t-value	df	p-value
Pooled Variance t-test:	-0.7619	58.0000	0.4491

You can request the system to provide the Nonparametric test by selecting the **Options** menu **Output** option. If you prefer, you can make your selections from the **System Preferences** menu options in either the SOLAS 3.0 Main window, or the Output window menus.



Two-group *t*-test

The two-group *t*-test checks for a specified difference in the population means using two independent samples that include an appropriate outcome variable. By default, the system tests the null hypothesis of equal population means "that the difference in the population means is zero". You can also test the null hypothesis that the difference in the population means is a non-zero value. If you are testing a non-zero value for the difference, type that value into the null hypothesis datafield.

When performing the two-sample *t*-test, you are assuming that you have normal data with equal variances in the two groups. If the variances are not equal (according to the results of Levene's test), there are three possibilities:

- You can use the separate variance *t*-test.

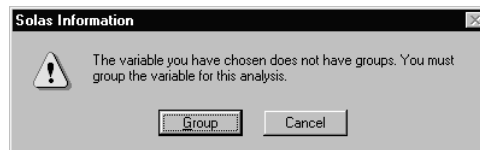
- Transform your data to provide a “closer to normal distribution” with approximately equal variance.
- You can use Nonparametric tests.

If the sample sizes are approximately equal, the two-sample t -test is quite robust.

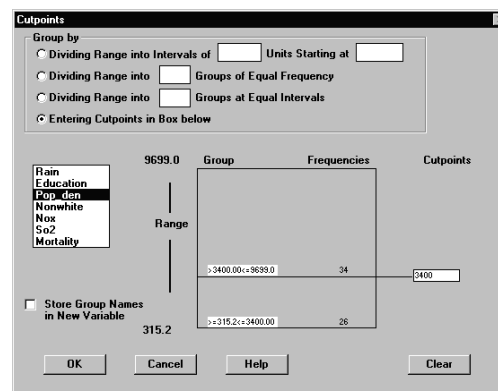
Although the data are obviously not ideal for t -tests, let us try using the air pollution data set - Airpoll2. mdd for this test. We will divide the population density into two groups, those cities with a density less than or equal to 3000, and those above 3000. For the Outcome variable, we will choose Mortality.

To perform a Two-group t -test:

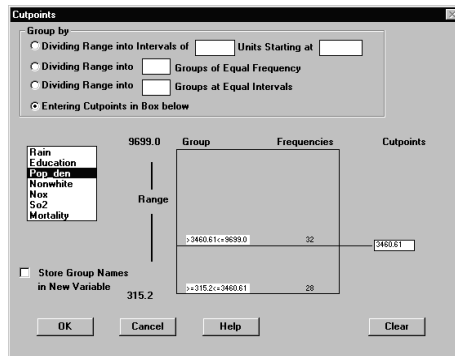
1. Drag and drop an Outcome variable and a Grouping variable to the appropriate datafields. A warning is displayed because Pop_den is a continuous variable. For this example, we have to set cutpoints for the Pop_den variable.



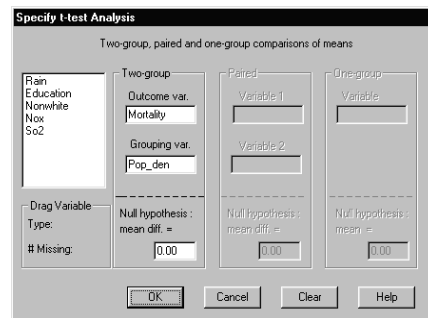
2. In the warning message window, select Group to group the variable Pop_den, and the Cutpoints window is displayed.
3. Check the “Entering cutpoints in box below” option, then click in the cutpoint graphics area to draw a cutpoint.
4. Click and hold the left mouse button on the cutpoint line, dragging the line up and down to get the cutpoint number 3400 in the display field at the right-hand side of the line.



5. If by dragging, you cannot get the cutpoint number to exactly 3400, just type the number 3400 in the cutpoint display field. The line then moves to the specified cutpoint number.



6. Click **OK** to enter the cutpoint and return to the Specify *t*-test window.



7. The Outcome (or response) variable is the variable being used in the test. The Grouping variable is the variable that specifies the group for each case. Use the **Clear** button if you want to remove variables from the datafields.
8. After entering the variables, click **OK**. The output is displayed in the Two-group *t*- and Non-param Tests Output window.

Two-group t and Non-param Tests : AIRPOLL2 - Mortality and Pop_den

File Edit View Options Format Plots Run Window Help

Arial 10 B I U

DESCRIPTIVE STATISTICS

	N	Mean	StdDev	StdErr
Mortality	60	940.3816	62.2124	8.0315
>=315.2 <=3513.03	30	934.2400	67.5727	12.3370
>3513.03 <=9699.0	30	946.5233	56.8340	10.3764
Diff. of Means		-12.2833	N/A	16.1205

Test Statistics :

	t-value	df	p-value
Pooled Variance t-test:	-0.7619	58.0000	0.4491

NUM Col: 0 Line: 64 Page: 1

Note that the higher density cities have a larger mean value than the lower density cities.

The output options that can be selected for this window include the t -test, descriptive statistics, Levene's test of equal variances, confidence intervals (default 95%), and the separate variance t -test. Additional output includes nonparametric statistics and robust statistics.

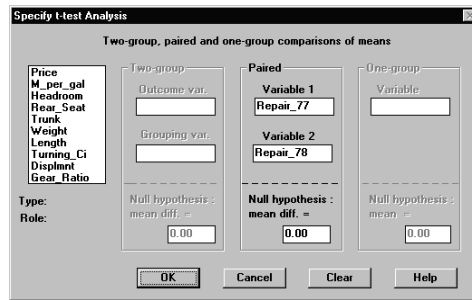
From an Output window, select **Options** → **Output** to display the Output Options window.

Paired t -test

In running this test, you are assuming that the differences between numerical values for the cases are normally distributed. If this assumption is not at least approximately true, you should consider transformations or Nonparametric tests.

By the nature of the design, the sample sizes of the two groups will be equal. The paired t -test requires two Outcome variables, but no Grouping variable. The Cars.mdd data set was used for the following example.

1. Drag and drop the desired variables from the Variables list to the relevant datafields. In this example, we have two Ordinal variables being treated as continuous. The system tests the null hypothesis that "the difference in the population means is zero". However, you can type in a non-zero value in the datafield.
2. After you click on **OK**, the Output window displays the results for the t -tests comparing two paired or dependent samples. You can choose additional output by selecting the **Options** menu **Output** option.

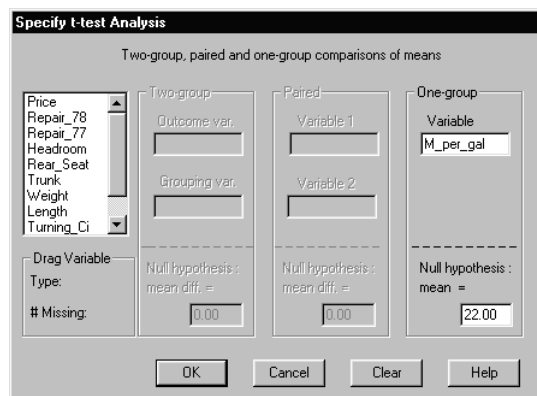


The Output Options for the *t*-test include the *t*-test, descriptive statistics, confidence intervals, nonparametric tests (sign test, Wilcoxon signed - rank test, and Spearman rank correlation) and trimmed mean *t*-test. These can be set in the Output Options window.

One-group *t*-Test

You use One-group tests when you want to test that the population mean is equal to a specified quantity. This test can be helpful in quality control applications. The specified quantity may be one that occurred in the past, or one that is expected by theory. The hypothesized value of the mean is usually not zero. Rather, the hypothesized value of the mean depends on the chosen variable. The output options include descriptive statistics, confidence limits on the mean, the *t*-test, the trimmed *t*-test, sign test, and the signed rank test.

When using the one-group *t*-test, you are assuming that the data are normally distributed. If this assumption is not approximately true, you should consider applying a transform, or using Non-parametric tests. The Cars.mdd data set was used for the following example:



1. Drag and drop your chosen Outcome variable to the datafield from the list of variables.
2. Enter the comparison mean value for the null hypothesis. You can use the **Clear** button to remove a variable from the datafield.
3. Press the **OK** button to display the Output window.

One-group t and Non-param Tests: CARS - M_per_gal

File Edit View Options Format Plots Berun Window Help

Arial 10 B I U

DESCRIPTIVE STATISTICS

	N	Mean	StdDev	StdErr
M_per_gal	74	21.2972	5.7855	0.6725

Test Statistics :

	t-value	df	p-value
t-test:	-1.0448	73.0000	0.2995

NUM Col: 0 Line: 48 Page: 1

Analysis of Variance

To perform an analysis of variance:

1. Start at the Datasheet window.
2. Click on the **Analyze** menu **ANOVA** option. The Specify ANOVA window is displayed.

The test setup is similar to that for the *t*- and Nonparametric Tests window. You enter the outcome variables and grouping variables in the same way as for the *t*-test (see preceding discussion).

One-way ANOVA

In one-way ANOVA, you are testing the null hypothesis that *k* independent population group means are equal, as opposed to the *t*-test in which you are testing that two means are equal. In this example, Mortality is the Outcome variable, and the variable Pop_den has been grouped into three groups.

Specify ANOVA

One-way ANOVA

Outcome var.
Mortality

Grouping var.
Pop_den

Two-way ANOVA

Outcome var.
Grouping var. 1
Grouping var. 2

Fit Additive model (no interaction)

Drag Variable

Type:
Missing:

OK Cancel Clear Help

The main ANOVA output is the analysis of variance table with the usual F statistic and its p -value. A small p -value signifies that the population means are likely to be different. In analysis of variance, you are assuming that the cases in the population are normally distributed and have equal variances. If you reject the null hypothesis, you may want to specify contrasts among the sample means.

Contrasts

To specify contrasts:

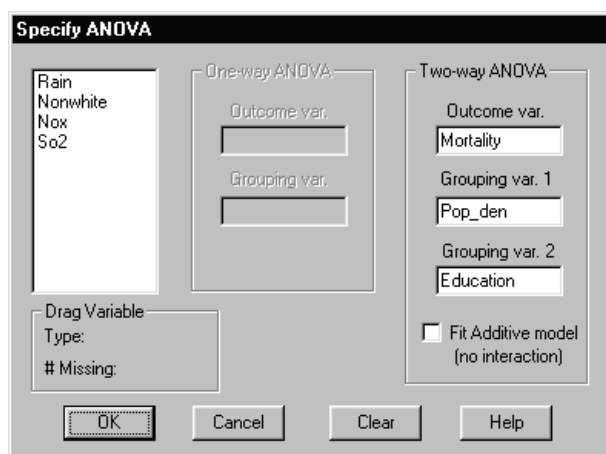
1. Click on the **Options** menu in the ANOVA Output window and choose **Contrasts**. Four options are available. The last two assume that the grouping variable is continuous (equal interval). Also see Chapter 6 - ANOVA in this manual.
2. Choose the desired option. You can also click on the **Options** menu and choose **Output**.

You can include the Kruskal-Wallis nonparametric test in your ANOVA output. This test uses the ranks of the data, assuming continuous data. The null hypothesis being tested is that k samples are drawn from k identical populations. If you have unequal variances, we recommend considering transforming the data. This is particularly recommended when the sample sizes are quite different.

Two-way ANOVA

In two-way analysis of variance, you need two grouping variables. The procedure is the same as that for one grouping variable. For this example, we have grouped the variable `Pop_den` into three groups (as for the One-way ANOVA previously) and the variable `Education` into two groups.

1. Drag the desired grouping variables into the Grouping fields and categorize as necessary.
2. Drag the outcome variable into the outcome field and click **OK**.



3. The ANOVA output window is displayed.

Two-way ANOVA : AIRPOLL2 - Mortality, Pop_den and Education

File Edit View Options Format Plots Run Window Help

Arial 10 B I U

DESCRIPTIVE STATISTICS

	N	Mean	StdDev
Mortality	60	940.3816	62.2124
>=315.2 <=3041.22	15	911.9333	62.3087
>3041.22 <=4509.08	30	947.2100	60.0266
>4509.08 <=9699.0	15	955.1733	61.3458
>=9.0000 <=10.86	26	970.7961	55.2651
>10.86 <=12.3000	34	917.1235	57.6178
>=315. , >=9.00	5	952.3599	47.8054
>=315. , >10.86	10	891.7200	60.5111
>3041. , >=9.00	13	971.1923	62.8901
>3041. , >10.86	17	928.8705	52.3426
>4509. , >=9.00	8	981.6749	49.4310
>4509. , >10.86	7	924.8857	62.6396

	StdErr	Min	Max
Mortality	8.0315	790.7000	1113.0000
>=315.2 <=3041.22	16.0880	790.7000	1018.0000
>3041.22 <=4509.08	10.9593	823.7999	1113.0000
>4509.08 <=9699.0	15.8394	861.4000	1071.0000
>=9.0000 <=10.86	10.8383	844.0999	1113.0000

NUM Col: 0 Line: 180 Page: 1

You can specify additional output by clicking on the **Options** menu **Output** option.

Regression Analysis

The X variable(s) is called the Independent or Predictor variable, and the Y variable is called the Dependent or Outcome variable. You should use the multiple regression option if you have more than one independent variable, and you may also use it when you have a single independent variable. If you prefer, you can use simple linear regression when you have only one independent variable.

To obtain simple linear regression or multiple regression results:

1. Click on the **Analyze** menu in the Datasheet window.
2. Choose **Regression**. You can use regression analysis to study the relationship between two variables, X and Y, or between a set of X variables and one Y variable.

Simple Linear Regression

1. Choosing **Simple Linear Regression** from the two **Regression** options displays the Specify Simple Regression window.

Specify Simple Regression

Y Variable (Dependent): Mortality

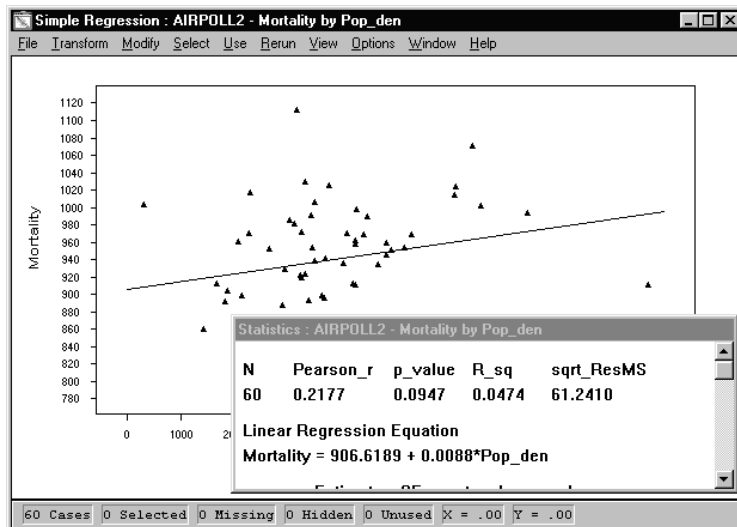
X Variable (Predictor): Pop_den

Equation: Mortality = Intercept + Slope * Pop_den + error

Drag Variable Type: # Missing:

OK Cancel Help

2. Drag the desired Y and X variables from the list of variables into the appropriate datafields, and click **OK**.



The output is a scatterplot with the Y variable on the vertical axis, and the X variable on the horizontal axis. A least squares regression line will be drawn on the scatterplot.

A Regression Statistics window will be in the lower right hand side of the screen. The regression statistics include:

- Sample size
- The Pearson product moment correlation and its square
- The p -value computed from a test of the null hypothesis that the population correlation is zero.

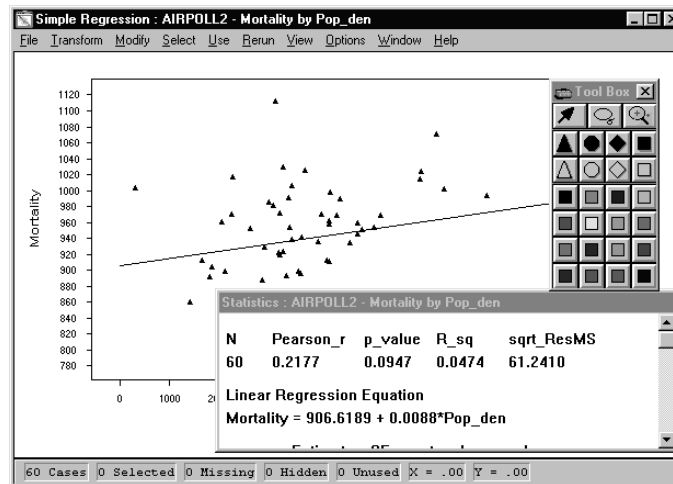
This p -value is the same value you get when you test that the population slope coefficient is zero.

In interpreting the regression model, it is useful to keep in mind the two basic models; the fixed effect model, for which the Xs are chosen at fixed levels; and the variable X model, for which multiple measurements are made on a single sample. With the exception of the interpretation of the squared multiple correlation, the correlations are not used in the fixed X models.

If the relationship between the variables does not appear to follow a straight line, you may want to transform either the X or the Y variable.

To transform a variable:

1. Click on a variable name on the Scatterplot X or Y axes.
2. Click on the Output window **Transform** menu, and choose an appropriate transformation. The Scatterplot changes to display the transformed variable. The statistics also change.
3. If there are outliers (points that are distant from the regression line), you can see the effect of removing them. Start by going to **View** menu, and choosing **Toolbox**.



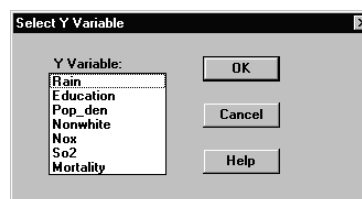
4. Click on the arrow at the top of the Toolbox and the arrow will be enabled. Then select a possible outlying point and click on it. The point will enlarge slightly, indicating that you have chosen it.
5. Click on the **Use** menu, and choose **Do Not Use Selected Points (Local)** if you just want to try out the effect of removing the chosen points. If you want the points to be grayed out in the datasheet and not used in other analyses, choose **Do Not Use Selected Points (Global)**.

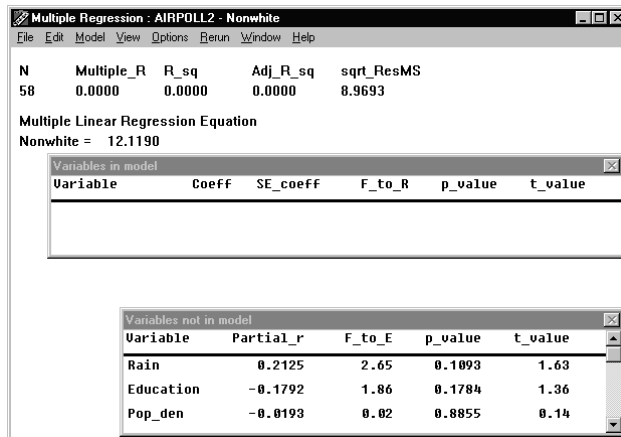
If you want to see a list of the additional statistics and options that are available, select the **View** menu. The **View** menu offers Descriptive Statistics, an ANOVA table (for testing that the population slope coefficient is zero), Plots (to check for Normality Probability and Residuals), and Diagnostics. The **View** option also includes additional outputs that you can add to the Scatterplot.

6. You'll find simple linear regression useful for visualizing the relationship between the Y variable and each X variable that you might use in a multiple regression. This is an easy-to-use screening tool for finding outliers, and roughly checking the linear model.
7. To close the Simple Regression window, click on the box in the upper right-hand corner of the screen.

Multiple Regression

1. For the Multiple Regression option, choose the Y variable and click the **OK** button to display the Output window.

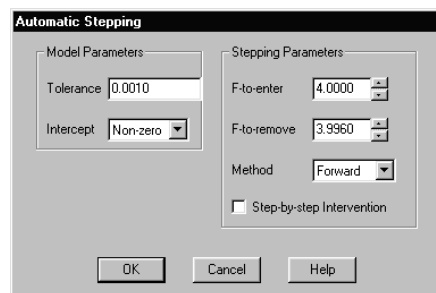




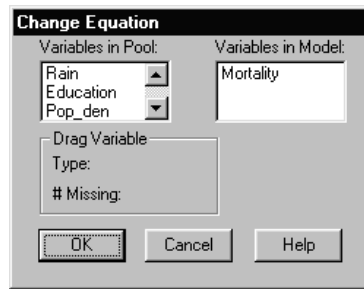
- You will see two windows. One window displays the variables in the model, and one window holds candidate X variables. You can enter each variable into the model by dragging the variable from the Variables not in model window to the Variables in model window. As soon as you drag a new variable into the model, the equation changes to reflect the presence of that variable.

You can modify the size of each of these windows by moving the cursor to an edge, holding down the mouse button and dragging the window edge until you reach the desired size.

If you prefer an automatic stepwise regression, you can click on the **Model** menu and choose the **Automatic Stepping** option. If you choose **Automatic Stepping**, you can select appropriate minimum *F*-to-enter and maximum *F*-to-remove values for forward stepwise regression. Make sure that your *F*-to-enter value is larger than your *F*-to-remove value. The system default value for tolerance is 0.001. A zero intercept can be specified if dictated by the model.



- After you set the parameters in the Automatic Stepping window, press **OK** to view the stepwise regression.
- If you choose **Change Equation** from the **Model** menu, you will see a list of the candidate X variables that you can drag to the Variables in model window. Click on the desired variables and click **OK**. The variables move to the Variables in model window.



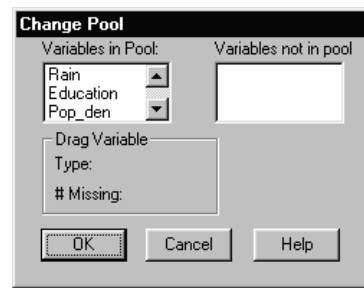
While interpreting the regression model, it is useful to keep in mind the two basic models:

- Fixed effect** The model in which the Xs are chosen at fixed levels.
- Variable X model** The model in which multiple measurements are made on a single sample.

With the exception of the interpretation of the squared multiple correlation, the correlations are not used in the fixed X models.

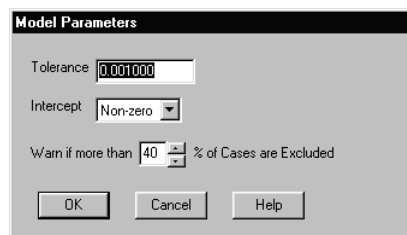
Change Pool

You can remove variables from the candidate pool by choosing the **Change Pool** option under the **Model** menu.



Change Model Parameters

The **Change Model Parameters** option allows you change the tolerance level, make an intercept zero, or change the percentage of excluded cases at which the system generates a warning message.



For more information about Simple Linear and Multiple Regression, see Chapter 3 – Regression in this manual.

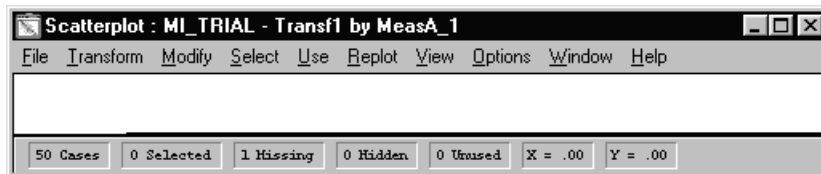
Plots

Six different types of output plot are available:

- Scatterplot, Histogram, Boxplot
- Barchart, Normal Probability Plot, Means Comparison Chart

All the available plots are described in Chapter 7 of this manual, and the “use of a Histogram in dealing with outliers” is described earlier in this tutorial. For the purpose of this section of the tutorial we will discuss the Scatterplot.

All of the functions associated with plots can be selected from the menus of a plot Output window:



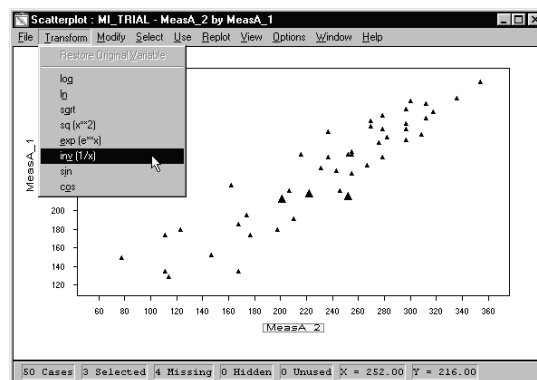
Each of the menus and their functionality are described in the following sections:

File menu

The **File** menu comprises the usual **Save**, **Print**, and **Close** functions, and an **Unlink** function that is described in Chapter 1 – “Data Management – Link manager”.

Transform menu

From the Transform menu the user can select from a choice of eight basic transform functions which can be directly applied to the plot variables. Highlighting a variable name in the plot output window enables the transform functions in the menu, with the exception of the **Undo Transform** function, which is only enabled immediately after applying a transform to a variable. After using the transformed variable, the function is disabled for that variable.



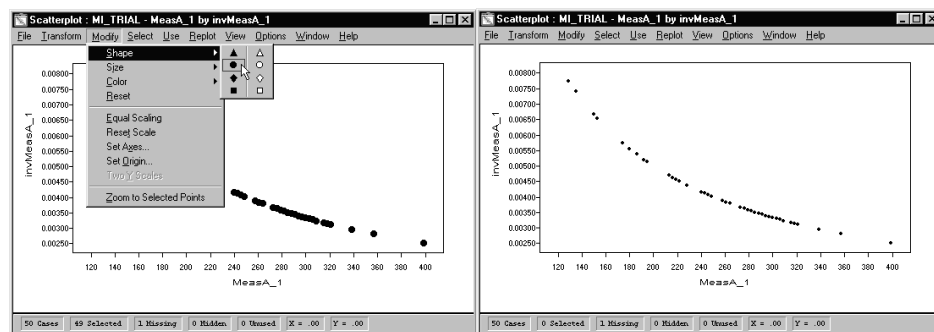
Highlighting a variable name, then selecting a transform function from the menu, displays the following message:



After transforming the selected variable, the **Undo Transform** function is enabled where you can undo the transform and delete the transformed variable from your datasheet. You can then apply a different transform function to the same variable. You can also make several different transforms of the same variable that can be stored in your datasheet. The **Undo Transform** function remains enabled for a variable until that transformed variable is used, after which the variable(s) can only be deleted from the datasheet.

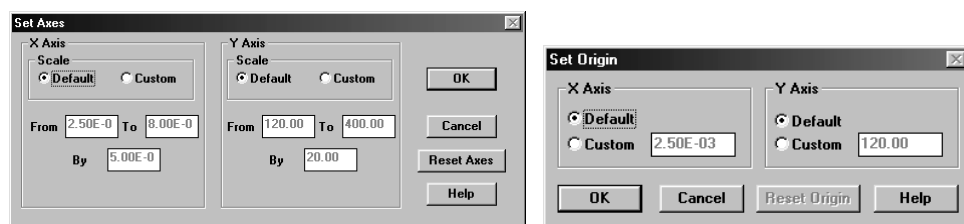
Modify menu

Using the functions in the **Modify** menu you can set the **Shape**, **Size** and **Color** of the points in a plot.



The picture above right shows a scatterplot of the transformed X variable - invMeasA_1 by Y variable - MeasA_1 (its non-transformed values). Note that the symbol shape preferences have been modified from triangular to round using functions in the **Modify** menu.

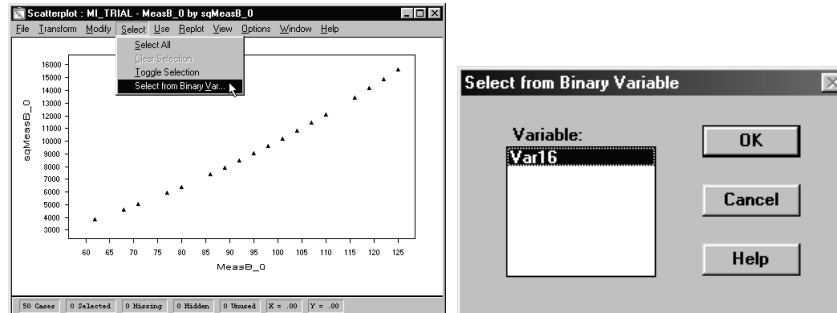
You can set the **Axes** and **Origin** of a plot using the windows shown below:



You can also **Zoom** and **Unzoom** to/from points in a plot that you have selected using the mouse pointer. There is also a way to select multiple points in a plot using the Toolbox, this is described later.

Select menu

Using the functions in the **Select** menu you can **Select All** points in a plot, invert an individual selection of points using the **Toggle Selection** function, and **Select from Binary_Var** to display a list of binary variables in your datasheet. If you have not assigned binary variables, this option is grayed out.



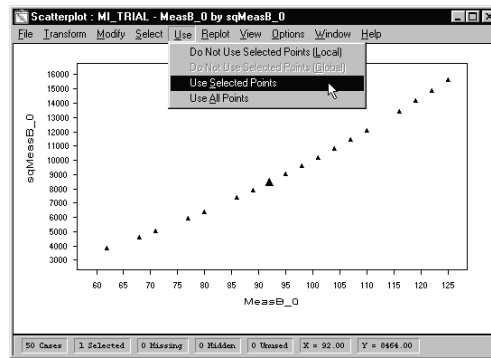
The system displays the Select from Binary Variable window. All Binary variables in your datasheet are displayed in the Variable listbox.

The value of each case of the binary variable determines whether the point representing that case is selected. Enlarged points, indicating that they are selected, represent all the cases for which the value of the binary variable is equal to 1. Normal-sized points, indicating that they are not selected, represent all the cases for which the value of the binary variable is equal to 0.

The Select from Binary Variable option lets you use any existing nominal variable with two categories to select points in the plot.

Use menu

The **Use** menu functions allow you select/de-select points in your plot. Points can either be de-selected locally (**Do Not Use Selected Points [Local]**) which affects only the selected points in the current plot, and highlights those points in the datasheet. When you use the **Do Not Use Selected Points [Global]** function, then all current results/plots/linked results from the same datasheet will be affected by the selection (not visible in plots), and the selected points/cases will be grayed-out in the datasheet.



Replot menu

From this menu you can use the **Specify New Scatterplot** function to employ different variables in the same design, or in a different design. You can also use the **Modify Current**

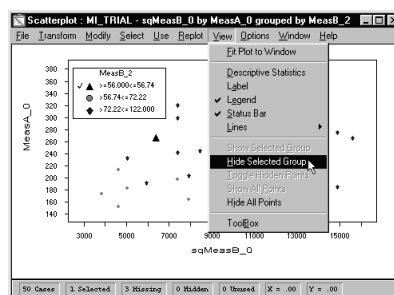
Scatterplot function to dynamically change the plot, observing the effects of using different variables on the output.

Another way of dynamically modifying a plot output is by using the Toolbox in conjunction with the **View** menu **Lines** option. This is described in the next section.

View menu

This menu has selections that allow you to view information relating to the points (cases) comprising the plot output and to display statistics for the plot. The following functions are available:

- Re-sizing a plot output window, displaying a legend panel, status bar and labels (case numbers).
- Showing/hiding points and groups, adding lines to a plot, displaying the descriptive statistics, and displaying the Toolbox.



The **Fit Plot to Window** option allows you to re-size the output window by moving the mouse cursor over an edge, or the lower right-hand corner, and dragging to the desired size. Then select this option to fit the plot into the re-sized window.

The **Descriptive Statistics** option displays a window showing the default statistics for a plot. Selecting the **Options** menu **Output** option in this window displays an Output Options window that allows you select additional statistics for display.

Checking the **Label** option displays the case number from the datasheet for the point(s) selected by the cursor, or from the Legend panel.

Checking the **Legend** option displays a panel showing the symbols used, the name of the grouping variable, and cutpoints for the groups. Clicking on a symbol selects (highlights) all the points associated with the selection in the plot.

Checking the **Status Bar** option displays the Status Bar at the lower end of the plot window. This shows the number of cases in a datasheet, how many points are selected in the plot window, how many cases are missing, how many points have been hidden, and how many cases are unused. Also the case values in the X and Y variables which form the co-ordinates of a selected point in the plot.

The **Lines** option allows you to place a **Linear Fit** line, and **Reference** lines for **Mean X** and **Mean Y** in the plot output. You can also generate a **Line plot**. After placing a line in a plot, you can apply different **Confidence Bands** in the range 0.90, 0.95 and 0.99.

The **Zero X** line and **Zero Y** line options are enabled when either axis contains a zero value.

The **y = x** **Line** option draws a line which has a slope of 1. The origin of the **y**-axis has to be in the range of the **x**-axis to enable this option.

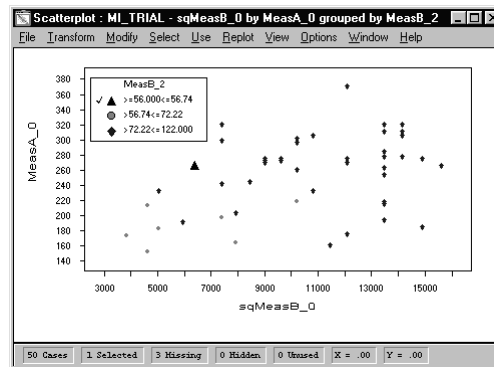
The **Descriptive Statistics** option displays default statistics that are set by the system. These results are for the Scatterplot shown below.

	N	Mean	StdDev	Variance
MeasA_0	47	253.9787	51.6187	2664.4995
>=56.000 <=78.12	10	210.6000	46.4930	2161.6000
>78.12 <=100.61	16	240.3750	44.2385	1957.0500
>100.61 <=122.000	21	285.0000	39.8559	1588.5000

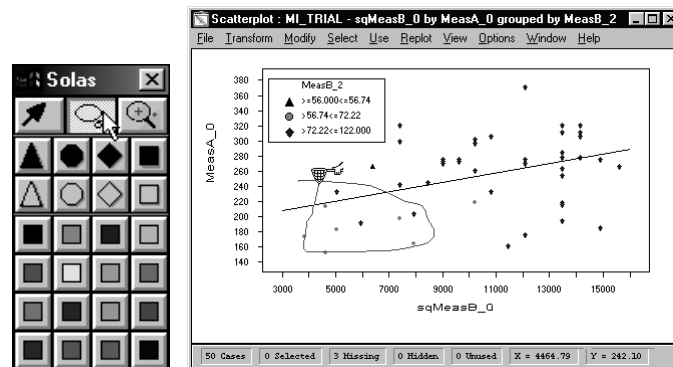
	Min	Max
MeasA_0	153.0000	372.0000
>=56.000 <=78.12	153.0000	300.0000
>78.12 <=100.61	162.0000	321.0000
>100.61 <=122.000	186.0000	372.0000

	N	Mean	StdDev	Variance
MeasA_1	47	245.4468	50.0370	2503.6568

You can select the **Options** menu **Output** option to display the Descriptive Statistics - Output Options window. Here you can select additional output results for display.

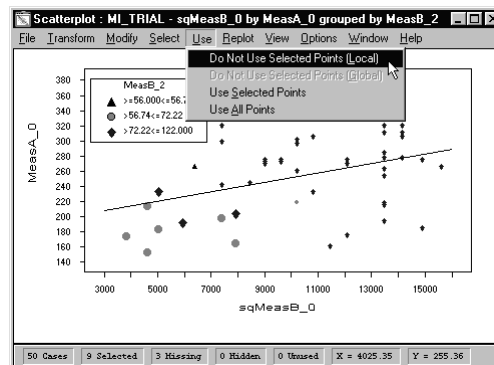


You can use the **Toolbox** option to select a point, or a group of points in a plot, zoom to points, and change the symbol shapes, and colors.

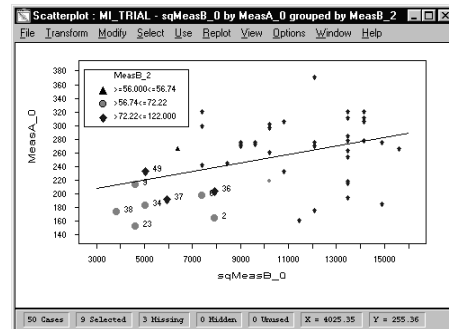


An example of using the Toolbox selection cursor is shown above. Simply click on the Select tool in the Toolbox (as shown), click and hold on your start point in the output window, then drag the cursor around your selected points to encircle them with a line.

Release the left mouse button to highlight the selected points (as shown below), then open the **Use** menu where you can select a “Do Not Use” function.



After selecting a function from the **Use** menu, the plot output changes accordingly as shown below:



You can revert to the original plot output by selecting the **Use All Points** function, then continue selecting different points to exclude. Note that the **Label** option has been selected to display the case numbers for the excluded points.

You can also select discrete points using the mouse cursor, or select groups of points by clicking on a symbol in the **Legend** panel.

Options menu

Selecting the **Statistics** option displays Linear Regression statistics window. The statistics displayed are the default calculations. Statistics for the above Scatterplot are shown below:

Statistics : MI_TRIAL - MeasA_0 by sqMeasB_0				
N	Pearson_r	p_value	R_sq	sqrt_ResMS
50	0.3539	0.0117	0.1252	52.5655
Linear Regression Equation				
MeasA_0 = 190.2264 + 0.0061*sqMeasB_0				
	Estimates	SE	t_value	p_value
Intercept	190.2264	25.9773	7.3228	2.3978E-09
Slope	0.0061	0.0023	2.6216	0.0117

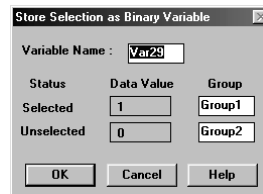
Selecting the **Diagnostics** option displays the Regression Diagnostics for your Scatterplot. The Diagnostics for the above Scatterplot are shown below:

Regression diagnostics: MI_TRIAL - MeasA_0						
	MeasA_0	Predicted	Residual	PI_Lower	PI_Upper	sqMeasB_0
	VVar					XVar
1	177	264.58	-87.58	157.61	371.55	12100.00
2	165	238.90	-73.90	131.41	346.40	7921.00
3	270	245.69	24.31	138.68	352.69	9025.00
4	276	249.25	26.75	142.40	356.09	9604.00
5	306	277.25	28.75	169.21	385.29	14161.00
6	198	235.68	-37.68	127.06	343.49	7396.00
7	147	272.92	-125.92	165.34	380.49	13456.00
8	321	272.92	48.08	165.34	380.49	13456.00
9	213	218.64	-5.64	108.22	329.06	4624.00
10	276	245.69	30.31	138.68	352.69	9025.00

The **Add Var to Datasheet** option has a menu with three selections that allow you to choose points in your Scatterplot, and store these as Group, or Binary Variables.

Store Selection as Binary Var

Selecting this option allows you to create a Binary Variable in the datasheet based on the selected points in the plot. Using the cursor or the Toolbox, simply select points in the plot, then select this option. You can rename the Binary Variable in the **Variable Name** datafield. The Binary Variable you create is appended to the datasheet.

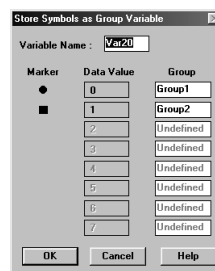


The selected items represent cases for which the Binary Variable will have value 1. The unselected item(s) in the plot represent cases for which the Binary Variable will have value 0.

Store Symbols as Group Var

Selecting this option allows you to create a Nominal Group Variable in the datasheet based on the selected symbols in the plot.

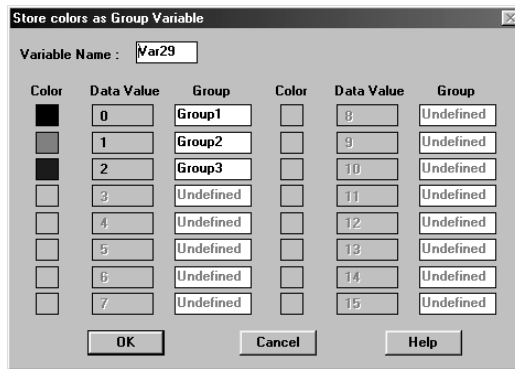
NOTE: If all the plot symbols are the same (distinguished only by color), you will need to change the symbol shape of the selected points, using the **Modify** menu, for this option to be enabled.



The system displays the Store Symbols as Group Variable window. Each field in the window is prefilled, based upon the group symbol shapes that are defined in the plot. The new Group Variable is given a default name, data values, and group names. You can edit the default names. The Group Variable that you create is appended to the datasheet.

Store Colors as Group Var

Selecting this option allows you to create a Nominal Group Variable in the datasheet based on the selected colors in the plot.



Each field in the Store Colors as Group Variable window is prefilled based on the colors defined in the plot. The new Group Variable with its default name, data values and group names is displayed. You can edit the default names by double clicking on a default variable or group name. The Group Variable that you create is appended to the datasheet.

Understanding the Link Manager

All of your results (analyses and plots) are linked to the datasheet (or frequency table) from which they were derived. Unless you break the link, a change in the datasheet is reflected in all of the results derived from it. Similarly, if you make a change to a result in an output window (regression or plot for example), the datasheet being used will reflect the change, and if the change is global, any other connected (linked) results will be affected. For example, suppose you add a case to a datasheet from which you have obtained three scatterplots, a multiple regression, and descriptive statistics. A new point will appear in the scatterplots, and both the regression and the descriptive statistics will be recomputed.

If you omit a point (globally) in one of the scatterplots, the case will disappear from all three scatterplots, and the case number will be grayed out in the datasheet (indicating that it is not in use). Finally, the regression and descriptive statistics will be recomputed without the omitted point. Your point selection is also reflected in all linked windows. If you highlight a set of points in one of the plots, the corresponding cases will be highlighted in the other plots and in the datasheet.

How the Link Manager Operates

The Link Manager comprises a powerful set of tools that you can use for screening data. Using the facilities provided by the Link Manager, you quickly demonstrate the effects of transforming your data, removing outliers, including or excluding variables, etc. Also, you can readily pick out related cases in various representations of the data. However, you need to understand how the Link Manager operates in order to use it to its full advantage.

The Link Manager maintains linkage between a datasheet, or frequency table, and all of its results. Certain kinds of changes automatically update all open linked windows immediately. These include adding or deleting cases, changing data values (including redefining a variable using a transformation), and changing a variable name.

Unless the change will invalidate open results, you will receive no warning about these updates. You may want to unlink certain results so that you can compare the results obtained before, and after a modification. Alternatively, you may prefer to copy the original datasheet and maintain two trees in your exploration of the data.

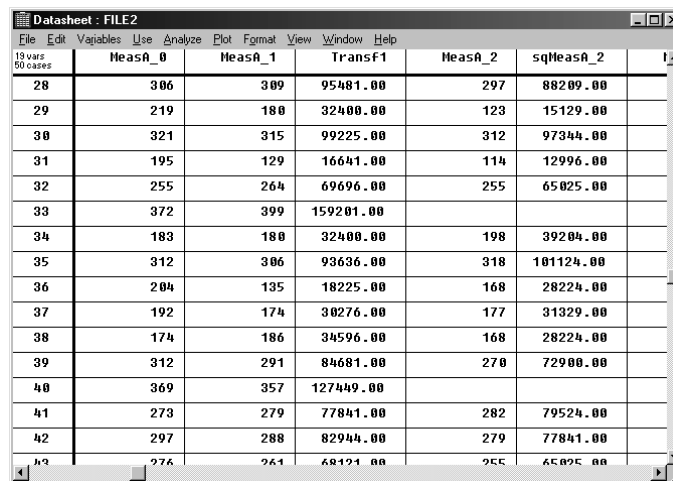
NOTE: Whether or not you unlink results, it is good practice to save a copy of your original data under a different name until you are sure that you no longer need it.

Examples showing how the Link manager operates in local and global modes are given below. For a more detailed explanation of the Link Manager functionality, see Chapter 1 – Data Management.

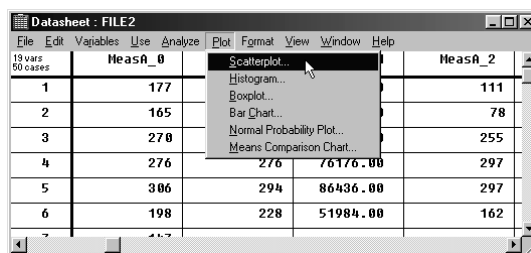
NOTE: Adding or deleting cases and variables, or modifying variables in a datasheet will affect ALL results for that datasheet. Thus changes to a datasheet are treated as global changes.

Using the Link manager in Local Mode - Example

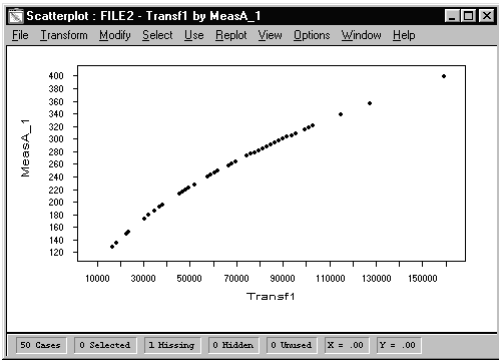
Using a copy of the datasheet **mi_trial.mdd** (FILE2) we transformed the variable MeasA_1 (Transf1 = $x**2$) and then generated a scatterplot:



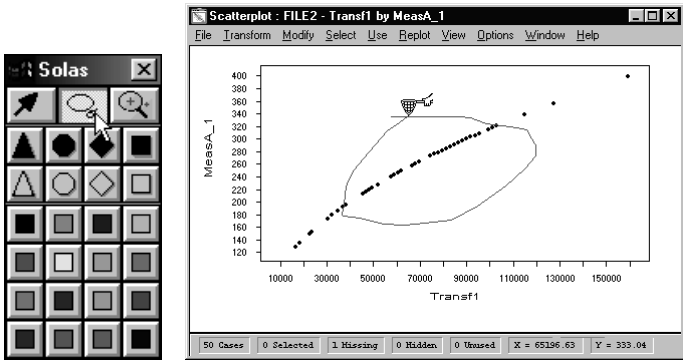
	MeasA_0	MeasA_1	Transf1	MeasA_2	sqMeasA_2
28	306	309	95481.00	297	88209.00
29	219	180	32400.00	123	15129.00
30	321	315	99225.00	312	97344.00
31	195	129	16641.00	114	12996.00
32	255	264	69696.00	255	65025.00
33	372	399	159201.00		
34	183	180	32400.00	198	39204.00
35	312	306	93636.00	318	101124.00
36	204	135	18225.00	168	28224.00
37	192	174	30276.00	177	31329.00
38	174	186	34596.00	168	28224.00
39	312	291	84681.00	270	72900.00
40	369	357	127449.00		
41	273	279	77841.00	282	79524.00
42	297	288	82944.00	279	77841.00
43	276	261	68121.00	255	65025.00



	MeasA_0	MeasA_1	Transf1	MeasA_2	sqMeasA_2
1	177			111	
2	165			78	
3	270			255	
4	276		76176.00	297	
5	306		93636.00	297	
6	198		39204.00	162	



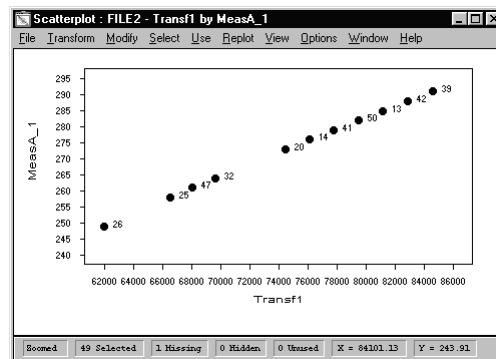
Selecting the Toolbox from the **View** menu, then choosing the Select tool, we select a number of cases from the plot that highlights the cases in the plot and the datasheet as shown below:



From the **Use** menu, we select the **Do Not Use Selected Points [Local]** option.

Datasheet : FILE2					
	MeasA_0	MeasA_1	Transf1	MeasA_2	sqMeasA_2
5	306	294	86436.00	297	88209.00
6	198	228	51984.00	162	26244.00
7	147				
8	321	321	103041.00	336	112896.00
9	213	213	45369.00	201	40401.00
10	276	216	46656.00	252	63504.00
11	285	288	82944.00	297	88209.00
12	303	303	91809.00	279	77841.00
13	273	285	81225.00	237	56169.00
14	279	276	76176.00		
15	186	192	36864.00	210	44100.00
16	300	297	88209.00	270	72900.00
17	243	222	49284.00	207	42849.00
18	216	222	49284.00	246	60516.00
19	279	258	66564.00	279	77841.00
20	270	273	74529.00	276	76176.00
21	321	300	90000.00	312	97344.00
22	267	261	68121.00	252	63504.00
23	153	153	23409.00	147	21609.00

Next we selected **Label** from the **View** menu of the plot output window, and also used the Zoom function in the Toolbox to help display some of the selected cases to be excluded from the analysis.



Selecting the **Options** menu **Statistics** item displays the Statistics window showing the modified Simple Linear Regression results for the remaining cases:

Statistics : FILE2 - MeasA_1 by Transf1				
N	Pearson_r	p_value	R_sq	sqrt_ResMS
30	0.9843	1.2080E-22	0.9689	11.4219
Linear Regression Equation				
MeasA_1 = 118.9952 + 0.0019*Transf1				
	Estimates	SE	t_value	p_value
Intercept	118.9952	4.0785	29.1761	1.6757E-22
Slope	0.0019	6.561E-05	29.5299	1.2080E-22

Selecting the **Use** menu, then **Use Selected Points**, displays a window showing the Simple Linear Regression results using all cases in the datasheet:

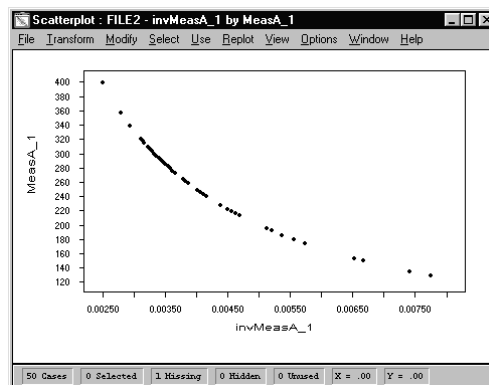
Statistics : FILE2 - MeasA_1 by Transf1				
N	Pearson_r	p_value	R_sq	sqrt_ResMS
49	0.9882	1.7520E-37	0.9765	9.4851
Linear Regression Equation				
MeasA_1 = 118.2967 + 0.0020*Transf1				
	Estimates	SE	t_value	p_value
Intercept	118.2967	3.2706	36.1694	1.7520E-37
Slope	0.0020	4.506E-05	44.1756	1.7520E-37

You can continue to select/deselect cases and choose output options either from a plot output window or the datasheet, at the same time monitoring the effects of including/excluding points/cases from your analysis.

NOTE: In “Local” mode, when you select points/cases, all open Regression/Plot outputs for the datasheet will show the highlighted points, but only the output results window from where the “Use Cases” options was selected will be subject to the modified analysis.

Using the Link manager in Global Mode – Example

Using a copy of the datasheet **mi_trial.mdd** (FILE2) we transformed the variable MeasA_1 ($\text{InvMeasA}_1 = \text{Inv}(1/x)$), generated a scatterplot, then selected some cases using the Toolbox Select tool, and from the **Use** menu, selected the **Do Not Use Selected Points [Global]** option to display the output shown below:

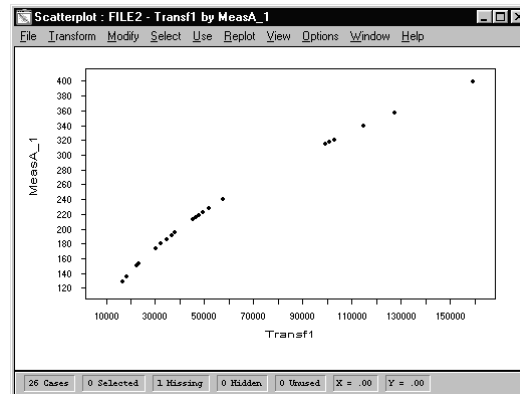


Notice that the selected cases are missing from the above plot, and that the same cases are “grayed out” in the datasheet shown below:

Datasheet: FILE2					
	MeasA_1	Var16	Var17	sqMeasA_1	invMeasA_1
8	192.00	1	1	103041.00	3.12E-03
9	297.00	1	2	45369.00	4.69E-03
10	321.00	1	2	46656.00	4.63E-03
11	213.00	1	1	82944.00	3.47E-03
12	216.00	1	1	91809.00	3.30E-03
13	288.00	1	1	81225.00	3.51E-03
14	303.00	1	1	76176.00	3.62E-03
15	285.00	1	2	36864.00	5.21E-03
16	276.00	1	1	88209.00	3.37E-03
17	192.00	1	2	49284.00	4.50E-03
18	297.00	1	2	49284.00	4.50E-03
19	321.00	1	2	66564.00	3.88E-03
20	213.00	1	1	74529.00	3.66E-03
21	216.00			90800.00	3.33E-03
22	288.00			68121.00	3.83E-03

Now, the “deselected” cases will be excluded from any linked or new analysis using this datasheet. Selecting the **Use All Cases** option in the datasheet **Use** menu will restore deselected case(s), but the datasheet and the linked results should be saved to a new file. In Global mode, the **Use All Cases** action can ONLY be performed from the datasheet, and NOT through the

Use menu of an output window. See “Recommended Methods for Observing the Effects of Modifications to your Data on your Output Results”.



Because the cases were deselected using the **Do Not Use Selected Cases [Global]** option (described above), they are also missing from another plot (shown above) that is linked to the datasheet **File2**.

Observing the Effects of Modifying your Data

The Statistical Solutions Systems for Data Analysis such as SOLASTM, and EquivTestTM allow you to demonstrate quickly the effects of experiments with your data on your output results. Changes such as:

- ◆ Transforming your data
- ◆ Removing outliers
- ◆ Including or excluding variables/cases
- ◆ Modifying variables

Using the following procedures enables you to easily pick out related cases in various representations of the data:

Modifying Plots and Simple Linear Regression Outputs

The functions in this procedure can be performed from a Plot or Regression output window, with the exception of **Use All Cases** which is selected from the datasheet **Use** menu.

1. Open a datasheet with Regression or Scatterplot results, and from the **File** menu select **Open Linked results**.
2. From the displayed window, select the result to be opened, then from the **File** menu in the result window select **Unlink** to display a Name for New Datasheet window.
3. Enter a name for the new datasheet and press the **OK** button. A new datasheet will be created with the variables used in the linked output result.
4. From the **File** menu in the new datasheet select **Save**, and from the **File** menu in the results window select **Save Result**.

You can then continue “unlinking” and “saving” results with different filenames creating a number of experimental datasheets with linked results. You can also copy variables from your original datasheet to the new datasheets at any time, or you can create new variables.

5. In the output results containing Scatterplots, you can select the **Options** menu **Statistics** function to display the Simple Linear Regression results in the Statistics window.
6. Open one of your saved results, select some points in the output using the mouse or the Selection tool in the Toolbox, then select **Label** from the **View** menu so you can easily identify the case numbers.
7. From the **Use** menu, select **Do Not Use Selected Points [Local]** (or **[Global]**).

You can see the effects of “not using” selected points/cases from the changes in the open Statistics window(s).

8. Using the **File** menu **Save As...** function, you can save the changed output results as “File n”,
- Or, from the **Use** menu in the datasheet, select **Use All Cases**, then repeat from Step 5 choosing different points in your output.

NOTE: If you chose the **Global** option in Step 6, and you choose to save the changed results as in Step 7, you must save the modified datasheet as a new file. Then, from the **File** menu in each modified result window, select **Save Result** to link these results with the modified datasheet.

9. You can repeat this procedure for all your experimental datasheets, selecting different points/cases, then saving the modified output as “File n+1” (as in Step 7) and so on.

Using the above procedure allows you to modify output results continuously, save the results in separate files and maintain the integrity of the links that are generated by the Link manager.

Selecting Variables/Cases and Applying Transforms from a Datasheet

Experiments with your output results can be performed exclusively from selections available on the datasheet menu-bar. Using these selections, the following types of changes can be applied:

- ◆ From the **Variables** menu you can change variable attributes, insert new variables, group variables, apply transformations, and define variables.
- ◆ From the **Use** menu you can choose which highlighted variables/cases to use, and define a systematic or a user-defined case selection algorithm.
- ◆ You can choose different analyses from the **Analyze** menu, or generate additional plots from the **Plots** menu.

Changing your output results by modifying the contents of a datasheet will apply those changes (globally) to all the output results linked to that datasheet, and because the results have been altered, they will be unlinked from that datasheet. To preserve your original datasheet, and any modified datasheets that you do not wish to save with their linked results, the following procedure is recommended:

1. Open your datasheet(s) containing linked results, and from the **File** menu select **Open Linked results**.
2. From the displayed window, select the result to be opened, then from the **File** menu in the result window select **Unlink** to display a Name for New Datasheet window.
3. Enter a name for the new datasheet and press the **OK** button. A new datasheet will be created with the variables used in the linked output result.
4. From the **File** menu in the new datasheet select **Save**, and from the **File** menu in the results window select **Save Result**.

You can then continue “unlinking” and “saving” results with different filenames creating a number of experimental datasheets with linked results. You can also copy variables from your original datasheet to the new datasheets at any time, or you can create new variables.

Close your original datasheets.

You can now open an experimental (source) datasheet and its linked results, and using functions from the datasheet menu-bar, begin applying changes (as described above) monitoring the effects of these changes on your output results.

When you are satisfied with your changes, save the datasheet as a new file, then link each changed result using the **Save Result** option in each result window **File** menu.

Or, if you are not satisfied with your changes, close the results windows, then close the datasheet window by selecting **No** in the SOLAS Information window. Then you can repeat from Step 6.

If you saved your results in Step 7, then from the source datasheet you can now select the **Use All Cases** Option from the **Use** menu to restore the original datasheet contents, and continue from Step 6 applying different conditions to its linked results.

Or, you can begin with a different experimental datasheet.

Using the above procedure you can continuously make significant modifications to your output results, save the results in separate files and maintain the integrity of the links that are generated by the Link manager.

Appendix A: Data Sets

DATA SETS

The system disks provide you with ten data sets for use in working with the program. This reference section gives you the origin for each of the data sets.

Airpoll2.mdd

The air pollution data, Airpoll.mdd, uses information from 60 U.S. metropolitan areas. For each record, the data include the following:-

Variables

1	Name	city name
2	Rain	mean annual precipitation in inches
3	Education	median school years completed for those over 25 in 1960 SMSA
4	Pop_den	population mile ² in urbanized area in 1960
5	Nonwhite	percentage of urban area population that is nonwhite
6	Nox	relative pollution potential of oxides of nitrogen, NO _x
7	So2	relative pollution potential of sulphur dioxide, SO ₂
8	Mortality	total age-adjusted mortality rate, expressed as deaths per 100,000

McDonald, G.C. and R.C. Schwing, 1973. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463-481.

Cars.mdd

The automobile data, Cars.mdd, are for cars sold in the United States during the 1979 model year.

Variables

1	Car	Make and model
2	Price	Price
3	M_per_gal	mileage
4	Repair_78	Repair record 1978
5	Repair_77	Repair record 1977
6	Headroom	Headroom in inches
7	Rear_Seat	Rear seat in inches
8	Trunk	Trunk space in cubic feet
9	Weight	Weight in pounds
10	Length	Length in inches
11	Turning_Ci	Turning circle in feet
12	Displmnt	Displacement in cubic metres
13	Gear_Ratio	Gear ratio

Chambers, J.M., W.S. Cleveland, B. Kleiner, and P.A. Tukey. 1983. *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

Fatness.mdd

The fatness data, Fatness.mdd, describe a group of 58 children born in Berkeley, California, between January 1928 and June 1929.

Variables

1	ID	identification number
2	Sex	identifies sex (coded 0 and 1)
3	wt_2	weight in kg at age 2
4	ht_2	height in cm at age 2
5	wt_9	weight in kg at age 9
6	ht_9	height in cm at age 9
7	leg_9	leg circumference at age 9
8	strong_9	strength measure at age 9
9	wt_18	weight in kg at age 18
10	ht_18	height in cm at age 18
11	leg_18	leg circumference at age 18
12	strong_18	strength measure at age 18
13	fatness	a measure of fatness on a 7-point scale, determined from a photograph at age 18:slender(1), fat(7)

Weisberg, S. 1980. *Applied Linear Regression*. New York:Wiley.

Fidell.mdd

The Fidell data, Fidell.mdd, are from a survey of 465 women by L. S. Fidell and J.E. Prather. The data consist of psychological and demographic measures, of which we use six (the names are changed slightly).

Variables

1	esteem	self-esteem measure coded from 1 to 22: low(1), high(22)
2	hap_stat	happiness with marital status coded from 1 to 48: low(1), high(48)
3	womenr ole	attitude toward role of women on scale from 1 to 38: conservative (1 -16), moderate(17 - 23), liberal(24 and up)
4	educatn	number of years of education
5	workstat	work status: paid work(0), homemaker(1 and 2)
6	marital	single(1), married(2)

Tabachnick, B.G. and L.S. Fidell. 1989. *Using Multivariate Statistics*. 2nd edition. New York: Harper and Row.

Health.mdd

The Health data, Health.mdd (containing 500 cases) are from a small simulation study conducted by Schafer (1997) to illustrate the frequentist properties of model-based multiple imputation when applied to a population of real data that do not conform to simple modeling assumptions.

He constructed an artificial population of 2000 subjects by drawing a simple random sample without replacement of the adult males in NHANES III who had complete data for the four variables in the table below:

Variables in the Simulation Study	
Variable	Description
AGE	Age group in years (1=20-39, 2=40-59, 3=over 60)
BMI	Body mass index (kg/m ²)
HYP	Hypertensive (1=no, 2=yes)
CHL	Total serum cholesterol (mg/dL)

Data for the simulation were taken from Phase 1 of the Third National Health and Nutrition Examination Survey (NHANES III) (National Center of Health Statistics, 1994).

Longley.mdd

The Longley data, Longley.mdd, were used by Longley to evaluate the accuracy of statistical algorithms in computer programs. For a complete listing of the implicit Price Deflators for Gross National Product, see Council of Economic Advisers, Economic Report of the President, January, 1964, Table C-6, p.214.

Variables

1	PriceDefla	GNP Implicit Price Deflator, 1954=100
2	GNP	Gross National Product
3	Unemployme	Unemployment
4	ArmForceSz	Size of armed forces
5	Population	Noninstitutional population 14 years of age and over
6	Year	Year
7	TotEmpl	Total derived employment
8	AgrEmp	Census agricultural employment
9	SelfEmp	Census self-employed
10	UnpFamWork	Census unpaid family workers
11	Domestics	Census domestics
12	BLSNonAgr	BLS nonagricultural private number of jobs
13	BLSFedGovt	BLS federal government
14	BLSLocGovt	BLS state and local government

Longley, J.W. 1967. An appraisal of least squares programs for the electronic computer from the point of view of the user, *Journal of the American Statistical Association*. 62:819-841.

Migrate.mdd

The migration data, Migrate.mdf, comes from a sample selected by the U.S. bureau of Census. The data compare region of residence in 1985 with 1980. Note that the migration data are frequency table data. If you want to open Migrate.mdf, you must select the Open dialog box to either .mdf files or to All files.

Categories

Northeast Residence in the Northeast

Midwest Residence in the Midwest

South Residence in the South

West Residence in the West

Agresti, A. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons.

Fisher.mdd and its altered version Fishmiss.mdd

The Fisher (1936) iris data, FISHER.mdd, contains measurements, in centimeters, of sepal length and width and petal length and width on samples of 50 irises from each of three species (1=Setosa, 2=Versicolor, 3=Virginica).

The file FISHMISS.mdd is a copy of the original file, but six values have been deleted at random.

Variables

1	Species	1=Setosa, 2=Versicolor, 3=Virginica
2	Sepallen	sepal length
3	Sepalwid	sepal width
4	Petallen	petal length
5	Petalwid	petal width

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179-184.

MI_TRIAL.mdd

The file MI_TRIAL.mdd is a simulated data set containing repeated measurements taken on two response variables MeasA and MeasB. The responses were measured at baseline, and again at month1, month2, and month3.

Variables

1	OBS	Observation number
2	SYMPDUR	The duration of symptoms
3	AGE	The patient's age
4	MeasA_0	The baseline measurement of the response variable MeasA
5	MeasA_1	The measurement of the response variable MeasA at month1
6	MeasA_2	The measurement of the response variable MeasA at month2
7	MeasA_3	The measurement of the response variable MeasA at month3
8	MeasB_0	The baseline measurement of the response variable MeasB
9	MeasB_1	The measurement of the response variable MeasB at month1
10	MeasB_2	The measurement of the response variable MeasB at month2
11.	MeasB_3	The measurement of the response variable MeasB at month3

Appendix B: References

REFERENCES

- Afifi, A.A. and V.A. Clark. 1990. *Computer-Aided Multivariate Analysis*. 2nd ed. New York: Van Nostrand Reinhold.
- Agresti, A. 1990. *Categorical Data Analysis*. New York: Wiley.
- Bendel, R.B. and A.A. Afifi. 1977. Comparison of Stopping Rules in Forward Stepwise Regression. *Journal of the American Statistical Association* 72:46-53.
- Bishop, Y.M.M., S.E. Feinberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- Box, G.E.P. and D.R. Cox. 1964. Analysis of transformations. *Journal of the Royal Statistical Society* 26:211-252
- Brown, M.B. and A.B. Forsythe. 1974. The small sample behaviour of some statistics which test the equality of several means. *Technometrics* 16:129-132.
- Chatterjee, S. And A.S. Hadi. 1989. *Sensitivity Analysis in Linear Regression*. New York: Wiley.
- Cook, R.D. and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Cramer, H. 1946. *Mathematical methods of Statistics*. Princeton: Princeton University Press.
- Daniel, C. And F.S. Wood. 1971. *Fitting Equations to Data*. New York: Wiley.
- Dixon, W.J. and F.J. Massey. 1983. *Introduction to Statistical Analysis*. 4th ed. New York: Wiley.
- Dunn, O.J. and V.A. Clark. 1987, *Applied Statistics: Analysis of Variance and Regression*. 2nd ed. New York:Wiley.
- Fleiss, J.L. 1981. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: Wiley.
- Forsythe, A.B., L. Engelman, R. Jennrich, and P.R.A. May. 1973. A stopping rule for variable selection in multiple regression. *Journal of the American Statistical Association* 68:75-77.
- Frigge, M., D.C. Hoaglin, and B. Iglewicz, 1989. Some Implementations of Box Plots. *The American Statistician* 43:50-54.
- Hoaglin, D.C., F. Mosteller, and J.W. Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.
- Kirk, R.E. 1982. *Experimental Design: Procedures for the Behavioural Sciences*. 2nd ed. Pacific Grove, CA:Brooks/Cole Publishing Co.

- Lehmann, E.L. 1975. *Nonparametrics: Statistical Methods based on Ranks*. Oakland, CA: Holden-Day.
- Mallows, C.L. 1973. Some comments on C_p . *Technometrics* 15:661-675.
- McDonald, G.C., and R.C. Schwing. 1973. Instabilities of regression estimates relating air pollution to mortality. *Technometrics* 15:463-481.
- Nishi, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* 12:758-765.
- Royston, J.B. 1982. An extension of Shapiro and Wilk test for normality to large samples. *Applied Statistics* 31:115-124.
- Shapiro, S.S. and M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52:591-611.
- Siegel, S. 1956. *Nonparametric Statistics for the Behavioural Sciences*. New York: McGraw-Hill.
- Velleman, P.F. and D.C. Hoaglin. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston MA: Duxbury Press.
- Velleman, P.F. and L. Wilkinson. 1993. Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician* 47:65-72.
- Welch, B.L. 1947. The generalisation of Students' problem when several different population variances are involved. *Biometrika* 34:28-35.
- Wickens, T.D. 1989. *Multiway Contingency Table Analysis for the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum Assoc.
- Winer, B.J., D.R. Brown, and K.M. Michels. 1991. *Statistical Principles in Experimental Design*. 3rd ed. New York: McGraw-Hill.
- Yates, F. 1934. Contingency tables involving small numbers and χ^2 test. *Journal of the Royal Statistical Society Supplement*. 1:217-235.
- Yuen, K.K. and W.J. Dixon. 1973. The approximate behaviour and performance of two-sample trimmed t. *Biometrika* 60: 369-74.

Appendix C: Error Messages

ERROR MESSAGES

1136	F_Eval Error: unimplemented builtin
1140	F_Eval Error: int function argument count error
1141	F_Eval Error: int function argument not number
1336	F_Eval Error: log function argument count error
1337	F_Eval Error: log function argument not number
1338	F_Eval Error: log of zero or negative number
1351	W_Eval Error: Invalid argument value
1352	W_Eval Error: Overflow range error
1353	W_Eval Error: Underflow range error
1354	W_Eval Error: Partial loss of significance
1355	W_Eval Error: Total loss of significance
1339	W_Eval Error: sin function argument count error
1340	W_Eval Error: sin function argument not numeric
1341	W_Eval Error: cos function argument count error
1342	W_Eval Error: cos function argument not numeric
1343	W_Eval Error: tan function argument count error
1344	W_Eval Error: tan function argument not numeric
1345	W_Eval Error: asin function argument count error
1346	W_Eval Error: asin function argument not numeric
1347	W_Eval Error: acos function argument not numeric
1348	W_Eval Error: acos function argument not numeric
1349	W_Eval Error: atan function argument count error
1350	W_Eval Error: atan function argument not numeric
1356	W_Eval Error: exp function argument count error
1357	W_Eval Error: exp function argument not numeric
1358	W_Eval Error: abs function argument count error
1359	W_Eval Error: abs function argument not numeric
1360	W_Eval Error: sign function argument count error

1361	W_Eval Error: sign function argument not numeric
1362	W_Eval Error: sqrt function argument count error
1363	W_Eval Error: sqrt function argument not numeric
1362	W_Eval Error: sqrt function argument count error
1363	W_Eval Error: sqrt function argument not numeric
1364	W_Eval Error: sq function argument count error
1365	W_Eval Error: sq function argument not numeric
1366	W_Eval Error: inv function argument count error
1367	W_Eval Error: inv function argument not numeric
1368	W_Eval Error: function argument count error
1369	W_Eval Error: function argument not numeric
1370	W_Eval Error: zero divisor
1371	W_Eval Error: function argument not integer
1373	W_Eval Error: invalid use of non-numeric variable
881	F_Lex Error: illegal src char
882	F_Lex Error: symbolic literal too long
883	F_Lex Error: eos in symbolic literal
884	F_Lex Error: illegal symbolic literal char
885	F_Lex Error: numeric literal too long
886	F_Lex Error: missing scale factor
887	F_Lex Error: missing scale factor sign
888	F_Lex Error: missing fractional part
889	F_Lex Error: name too long
890	F_Lex Error: eos in string
891	F_Lex Error: string too long
892	F_Lex Error: illegal escape sequence in string
893	F_Lex Error: empty symbolic literal
894	F_Lex Error: eos in alternate input
895	F_Lex Error: illegally formed ellipsis or invalid period
896	W_Parse Error: assignment operator expected
898	W_Parse Error: right parenthesis expected
903	W_Parse Error: loop keyword expected
904	W_Parse Error: function name expected
905	W_Parse Error: parameter name expected
906	W_Parse Error: variable name expected
909	W_Parse Error: right bracket expected
910	W_Parse Error: primary expected

912	W_Parse Error: name expected
918	W_Parse Error: invalid parameter
922	W_Parse Error: illegal parameter type
1089	W_Parse Error: illegal assignment statement
953	F_value Error: compare and type mismatch
954	F_value Error: type cannot be compare and
955	F_value Error: attempt to assign function
956	F_value Error:can't read data set element
959	F_value Error: illegal type for data set element assignment
961	F_value Error: non-boolean assignment to boolean element
962	F_value Error: non-numeric assignemnt to numeric element
965	F_value Error: non-string assignment to string element
966	F_value Error: undefined name before subscript
969	F_value Error: too many arguments
970	F_value Error: argument keyword is not parameter
971	F_value Error: argument cannot be function
999	F_value Error: non-numeric assignment ot numeric component
1000	F_value Error: non-boolean assignment to boolean component
1001	F_value Error: non-string assignment to string component
941	F_value Error: boolean value expected
942	F_value Error: assignment target not defined
943	F_value Error: numeric value expected
944	F_value Error: zero divisor
945	F_value Error: root of negative number
946	F_value Error: runtime error tap
947	F_value Error: undefined name in expression
948	F_value Error: undefined name in function call
949	F_value Error: name in call not function
1090	F_value Error: missing condition on case#
1091	F_value Error: illegal use of MISSING keyword
1092	F_value Error: too many names in expression
1093	F_value Error: argument out of range
861	Import Error: An unexpected end of file has been encountered
862	Import Error: file has a content other than DATA.

Appendix D: External File Formats

EXTERNAL FILE FORMATS

External Data Types

When importing or exporting files the following rules about data types are observed:

Numeric data are treated as Continuous if the absolute value is within the range 1.7e-308 and 1.7e+308. Numbers outside this range will be treated as missing.

Date data (when explicitly formatted as such) are treated as Character but can be converted to a numeric format by counting the number of days from 01 January 1900 and treated as Continuous.

String data are treated as Character. String data in a numeric variable is treated as missing data.

However, when importing, it is possible to treat any of the three data formats, (**Numeric**, **Date**, and **String**) as Continuous or Character using the Variable Attributes window in the datasheet **Variables** menu.

1 - 2 - 3 Worksheet Files

The system will read and write files from most versions of Lotus 1-2-3 (including Release 3.x and Windows).

Standard Extensions:

Lotus 1-2-3 Version 1 through 2.x	WK1
Lotus 1-2-3 Version 3.x and Windows	WK3

Reading 1-2-3 Worksheet Files

Because worksheet files are in general not designed to hold statistical data, only worksheets in certain formats can be read.

Worksheets must be in worksheet database format or in certain modifications of this. It is most straightforward with worksheets in database format.

Database format

Worksheet database files are structured worksheets where each row is a single case and each column contains a variable.

The first non-blank row of a worksheet database file has labels in each column that give the names of the variables. The data then begins in the next row. Data can consist of numbers (including serial date numbers), labels, or formulae.

After identifying the label row, the system will look up to 64 data rows to determine the data type of each column. If the first non-empty data cell of a particular

column is a number (or a label with a single period), the system will transfer the column as a number. If the data cell contains a label, the variable will be transferred as a string.

The width of the column for each numeric variable and the format of the first non-blank data cell in that column are used, where possible, to set the default target, or output, types for the numeric variables. Any date format in the first data row will set the target type to 'date'.

The column widths are used to set the maximum width of character variables, so be sure that columns are set wide enough to display all instances of each character variable.

The system is lenient in typing variables from worksheets. If it is expecting a character variable and it encounters a number it will convert it to a string.

Variations on Database Format

It is usually possible for the system to read a worksheet that has only data columns, without an initial row defining the variable names.

The system will look at the first two non-blank rows and will use them to detect whether the first row should be treated as labels or not. If there is any column which changes from a label in the first non-blank row to a number in the first non-blank row to a number in the second, the worksheet will be assumed to be in database format and the first non-blank row will be used as variable names. If there is no such type change, the system will treat the whole worksheet as data and assign variable names of the form 'Cn', where n is the column number.

Writing 1-2-3 Worksheet Files

On output, the system will write variable labels in the first row of the worksheet. Data values will be placed in the second and succeeding rows.

Column widths and formats will be determined by the variable information available. Dates and character variables are straightforward. For numerical data, information on the width and number of decimal places of variables, where available, is used to set the column widths and formats.

Missing Data

When importing data, blank cells are represented by the default missing value symbol. When transferring data to worksheets missing values will be written out as blank cells.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a Lotus 1-2-3 worksheet file are shown below in the first table. The second table shows the actual Lotus 1-2-3

output variable type that results from each the system target variable type assigned during a transfer from some other format.

Lotus 1-2-3 Input to the System

Contents of First Data Row	Default Target Type*
Label (only a period)	Continuous
Label (any other contents)	Character (length determined by column width)
Date format	Character
Numeric format in cell	Continuous

*See Overriding Default Target Type Section

Output to Lotus 1-2-3 from the System

Target Type	Output
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Ordinal Nominal	Label

dBASE Files and Compatibles

The system will read and write dBASE III+ and IV files, and those from compatible systems such as Clipper or Alpha Four. Obsolete dBASE II files can also be read.

Standard Extension: **DBF**

All versions of dBASE files can have indices for key fields, which are stored in separate fields. The system ignores these indices, and treats all files sequentially.

On input, dBASE numeric data and character variables are converted in a straightforward manner. Logical variables are converted to numbers ('True' becomes '1', 'False' becomes '0'). Memo fields cannot be covered and will not appear on the variable selection menu. Deleted records are not transferred.

Writing dBase Files

Users should be aware that dBASE fields are limited to 128 variables.

dBASE stores numeric data in fixed length character format. It is thus not very suitable for numbers which vary widely in magnitude or which are either very large or very small.

When the system is transferring data from a system in which the width and number of decimal places are known, it uses that information to set the format of each field in the output dBASE files. For systems, such as SYSTAT, in which this information is not recorded in the file, the system sets the formats based on the target type of the variable.

Missing Data

dBASE does not directly support missing values. On input to the system, blanks in a dBASE file are interpreted as missing values. If a data set is being transferred to a dBASE format, missing values in the input files are set to blank in the dBASE file. Blanks are interpreted as zero by dBASE. Many other programs, including the system, interpret these blanks as missing.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a dBASE file are shown below in the first table. The second table shows the actual dBASE output variable type that results from each system target variable type assigned during a transfer from some other format.

dBASE Input to the System

Input type	Default Target Type*
Numeric	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character	Character
Logical	Continuous
Date	Character
Memo	not translated

*See Overriding Default Target Type Section

Output to dBASE from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

Excel Worksheets

The system will read and write files from Excel. It will read all versions but will write only Version 2.1 files that can be directly and transparently read by any version.

Standard Extension: XLS

Reading Excel Worksheet Files

Because worksheet files are in general not designed to hold statistical data, only worksheets in certain formats can be read.

Worksheets must be in worksheet database format or in certain modifications of this. It is most straightforward with worksheets in database format.

Database format

Worksheet database fields are structured worksheets where each row is a single case and each column contains a variable.

The first non-blank row of a worksheet database file has labels in each column that gives the names of the variables. The data then begins in the next row. Data can consist of numbers (including serial date numbers), labels or formulas.

After identifying the label row, the system will look at up to 64 data rows to determine the data type of each column. If the first non-empty data cell of a particular column is a number (or a label with a single period), the system will transfer the column as a number. If the data cell contains a label, the variable will be transferred as a string.

The width of the column for each numeric variable and the format of the first non-blank data cell in that column are used, where possible, to set the default target, or output, types for the numeric variables. Any date format in the first data row will set the target type to 'date'.

The column widths are used to set the maximum width of character variables, so be sure that columns are set wide enough to display all instances of each of each character variable.

The system is lenient in typing variables from worksheets. If it is expecting a character variable and it encounters a number it will convert it to a string.

Variations on Database Format

It is usually possible for the system to read a worksheet that has only data columns, without an initial row defining the variable names.

The system will look at the first two non-blank rows and will use them to detect whether the first row should be treated as labels or not. If there is any column which changes from a label in the first non-blank row to a number in the second, the worksheet will be assumed to be in database format and the first non-blank row will be used as variable names. If there is not such a type change, the system will treat the whole worksheet as data and assign variable names of the form 'Cn', where n is the column number.

Writing Excel Worksheet Files

On output, the system will write variable labels in the first row of the worksheet. Data values will be placed in the second and succeeding rows.

Column widths and formats will be determined by the variable information available. Dates and character variables are straightforward. For numerical data, information on the width and number of decimal places of variables, where available, is used to set the column widths and formats.

Missing Data

When importing data, blank cells are represented by the default missing value symbol. When transferring data to worksheets missing values will be written out as blank cells.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in an Excel worksheet file are shown below in the first table. The second table shows the actual Excel output variable type that results from each the system target variable type assigned during a transfer from some other format.

Excel Input to The System

Contents of First Data Row	Default Target Type*
Label (only a period)	Continuous
Label (any other contents)	Character (length determined by column width)
Date format	Character
Numeric format in cell	Continuous

*See Overriding Default Target Type Section

Output to Excel from The System

Target Type	Output
Continuous	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Ordinal	
Nominal	
Integer	
Character	Label
Ordinal	
Nominal	

FoxPro Files

The system will read and write FoxPro Files.

Standard Extension: **DBF**

Reading FoxPro Files

FoxPro files can have indices for key fields, which are stored in separate files. The system ignores these indices, and treats all files sequentially.

On input, FoxPro numeric date and character variables are converted in a straightforward manner. Logical variables are converted to numbers ('True' becomes '1', 'False' becomes '0'). Memo fields cannot be converted and will not appear on the variable selection menu. Deleted records are not transferred.

Writing FoxPro Files

Users should be aware that FoxPro fields are limited to 128 variables.

FoxPro stores numeric data in fixed length character format. It is thus not very suitable for numbers which vary widely in magnitude or which are either very large or very small.

When the system is transferring data from a system in which the width and number of decimal places are known, it uses that information to set the format of each field in the output FoxPro files. For systems, such as SYSTAT, in which this information is not recorded in the file, the system uses sets the formats based on the target type of the variable.

Missing Data

FoxPro does not directly support missing values. On input to the system, blanks in a FoxPro file are interpreted as missing values. If a data set is being transferred to an FoxPro file format, missing values in the input files are set to blank in the FoxPro files. Blanks are interpreted as zero by FoxPro. Many other programs, including the system, interpret these blanks as missing.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in an FoxPro field are shown below in the first table. The second table shows the actual FoxPro output variable type that results for each The system target variable type assigned during a transfer from some other format.

FoxPro Input to the System

Input Type	Default Target Type*
Numeric	Continuous
Character	Character
Logical	Continuous
Date	Character
Memo	not translated

*See Overriding Default Target Type Section

Output to FoxPro from the System

Target Type	Output Type
Continuous	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Ordinal	
Nominal	
Integer	
Character	Character
Nominal	

Gauss Files

The system will read and write Gauss data sets. There are two Gauss formats. The first, Gauss 89, is used on PC platforms, and consists of two fields: a data file with a .DAT extension and a header, or dictionary file with a .DHT extension. The second Gauss format, Gauss 96, is used on UNIX platforms and has a single file with a .DAT extension.

Standard Extension: **DAT**

Reading Gauss Files

When you wish to transfer data from a Gauss data set, give the system the name of the data file (the file with the .DAT extension). If the system can find the .DHT file in the same directory, it will read the data file as a Gauss 89 file. If no .DHT files is present, the data file will be read as a Gauss 96 file.

Writing Gauss Files

On output, you can choose whether to write a Gauss 89 or Gauss 96 field. If you choose to write a Gauss 89 file, both of the Gauss files, the data file and the header file, will be written. The system will show the data file name, with the .DAT extension, the header file will be created as well, with a .DHT extension.

Missing Data

Gauss supports missing values.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a Gauss file are shown below in the first table. the second table shows the actual Gauss output variable type that results from each the system target variable type assigned during a transfer from some other format.

Gauss Input to the System

Input Type	Default Target Type*
Number	Continuous
Character	Character

*See Overriding Default Target Type Section

Output to Gauss from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character (8 byte maximum)

Minitab

Standard Extension: MTW

Reading Minitab Files

Minitab variable names can be up to 15 characters in length.

Versions 8 to 12 can be read by the system.

Writing minitab Files

Version 11 can be written by the system.

Any variable name in the source data set containing a left-parentheses followed by a number will be transferred into a MINITAB subscripted variable.

Users should note that the MINITAB error message, “You are trying to read an empty file”, will occur when MINITAB cannot find a data file. Your MINITAB files should be in the default drive or directory.

NOTE: Multiple worksheets stored in one file are not supported.

Missing Data

MINITAB supports missing values.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a MINITAB file are shown below in the first table. The second table shows the actual MINITAB output variable type that results from each the system target variable type assigned during a transfer from some other format.

MINITAB Input to the System

Input Type	Default Target Type
Numbers	Continuous
Character	Character

Output to MINITAB from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

Paradox Tables

Because Paradox stores numbers in binary rather than character representation and because it explicitly supports missing values, it is a much more suitable file format for statistical data than the dBASE format.

Standard Extension: DB

Reading Paradox Files

Paradox variable names can be up to 25 characters in length.

Paradox's date format is supported on input.

Writing Paradox Files

The system stores numbers into Paradox's integer format if they will fit. Paradox's date format is supported on output.

Missing Data

Paradox supports missing values for all data types.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a Paradox file are shown below in the first table. The second table shows the actual Paradox output variable type that results from each system target variable type assigned during a transfer from some other format.

Paradox Input to the System

Input Type	Default Target Type*
Numeric Dollar #(BCD)	Continuous
Short	Continuous
Long Autoincrement	Continuous
Alphanumeric	Character
Date Timestamp	Character

*See Overriding Default Target Type Section

Output to Paradox from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Ordinal Nominal	Alphanumeric

Quattro Pro Worksheet Files

The system will read and write Quattro Pro fields. QuattroPro version 8 is read-only.

Standard Extension: WQ*, WB*

Reading Quattro Worksheet Files

Because worksheet files are in general not designed to hold statistical data, only worksheets in certain formats can be read.

Worksheets must be in worksheet database format or in certain modifications of this. It is most straightforward with worksheets in database format.

Database format

Worksheet database files are structured worksheets where each row is a single case and each column contains a variable.

The first non-blank row of a worksheet database file has labels in each column that give the names of the variables. The data then begins in the next row. Data can consist of numbers (including serial date numbers), labels or formulas.

After identifying the label row, The system will look up to 64 data rows to determine the data type of each column. If the first non-empty data cell of a particular column is a number (or a label with a single period), The system will transfer the column as a number. If the data cell contains a label, the variable will be transferred as a string.

The width of the column for each numeric variable and the format of the first non-blank data cell in that column are used, where possible, to set the default target, or output, types for the numeric variables. Any date format in the first data row will set the target type to 'date'. The column widths are used to set the maximum width of character variables, so be sure that columns are set wide enough to display all instances of each character variable.

The system is lenient in typing variables from worksheets. If it is expecting a character variable and it encounters a number it will convert it to a string.

Variations on Database Format

It is usually possible for the system to read a worksheet that has only data columns, without an initial row defining the variable names.

The system will look at the first two non-blank rows and will use them to detect whether the first row should be treated as labels or not. If there is any column which changes from a label in the first non-blank row to a number in the second, the worksheet will be assumed to be in database format and the first non-blank row will be used as variable names. If there is not such a type change, The system will treat the whole worksheet as data and assign variable names of the form 'Cn', where *n* is the column number.

Writing Quattro Worksheet Files

On output, the system will write variable labels in the first row of the worksheet. Data values will be placed in the second and succeeding rows.

Column widths and formats will be determined by the variable information available. Dates and character variables are straightforward. For numerical data, information on the width of and number of decimal places of variables, where available, is used to set the column widths and formats.

Missing Data

When importing data, blank cells are represented by the default missing value symbol. When transferring data to worksheets missing values will be written out as blank cells.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a Quattro worksheet file are shown below in the first table. The second table shows the actual Quattro output variable type that results from each The system target variable type assigned during a transfer from some other format.

Quattro Pro Input to the System

Contents of First Data Row	Default Target Type*
Number (no format)	Continuous
Label (only a period)	Continuous
Label (any other contents)	Character (length determined by column width)
Date format	Character

*See Overriding Default Target Type Section

Output to Quattro Pro from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Ordinal Nominal	Label

SAS Data Files

The system will read and write SAS data files for the following platforms:

Windows & OS/2

SAS JMP including version 3

Standard extension: SD2

UNIX HP/Sun/IBM

Standard extension: SSD01

UNIX DEC

Standard extension: SSD04

Reading SAS Data Files

The system will read data files written by SAS version 6.08 and above.

SAS data files differ significantly between platforms. The system will read files written by SAS for Windows and OS/2, and files written for UNIX on HP, Sun, IBM and DEC Alpha platforms.

The system will automatically recognise the type of SAS file on input.

If you are moving SAS data files between platforms, you should be sure that you use a binary file transfer method.

Variable labels are supported, but unfortunately, value labels are not stored as a part of the SAS data file and can therefore not be transferred.

Writing SAS Data Files

On output, The system will let you choose one of the three supported file types listed above.

When writing SAS data files, you should pick an output format that is appropriate for the version of SAS that will be reading the file.

Missing Data

SAS supports missing values. On input, all of SAS's missing values are converted to a single internal missing value in the system. On output to SAS, missing values are set to '.', the SAS standard missing value.

Input and Output Variable Types

The target output variable types selected by the system for each input variable in a SAS data file are shown below in the first table. The second table shows the actual SAS data file output variable type that results from each the system target variable type assigned during a transfer from some other format.

SAS Data File Input to the System

Input Type	Default Target Type*
Numeric	Continuous unless print format is: DATE DDMMYY MMDDYY MMYY WORD, or WEEK, in which case it becomes a Character
Character Nominal	Character

*See Overriding Default Target Type Section

SAS Transport Files

The system will read and write data sets in the SAS Transport Format. This is, according to the SAS Institute, the best overall format for interfacing with other systems because it is consistent across all host environments. If you are downloading or uploading SAS data between computers, be sure to use an error-correcting file transfer protocol that is suitable for binary fields.

Standard Extension: **XPT**

Working with Transport Files within SAS

The method for writing (and reading) transport fields within SAS unfortunately varies across versions of SAS. For release 6.x users, the file can be written by any DATA or PROC step that creates SAS data sets and, similarly, it can be read by any DATA or PROC step. Most commonly, PROC COPY

is used to write (or to read) transport data sets. Release 6.03 users should use the SASV5XPT engine by naming it on the LIBNAME statement that defines the libref for the transport file. If you are using Version 6.06 or higher, you can read or write transport files by using the XPORT engine. To do so, you must name the XPORT engine in the LIBNAME statement.

For example, in Version 6.06 and higher the following code will write a transport file:

```
/* read system file 'old' - write transport file 'trans' */
libname old file-specification;
libname trans xport file-specification;
proc copy in=old out=trans;
run;
```

The resulting transport file can then be used for a the system data transfer.

If a transport file has been produced by the system, it can be read in SAS with the following.

```
/* read transport file 'trans' - write system file 'new' */
libname trans xport file-specification;
libname new file-specification;
proc copy in=trans out=new;
run;
```

Version 5 users can write and read transport files by using the XCOPY procedure, the COPY procedure with the EXPORT option, or the TRANSPORT=data set option.

For further information on how to read and write transport files, particularly using Version 5, see SAS Technical Report P-195, Transporting SAS Files between Host Systems and the documentation for SAS on your operating system. We also recommend, particularly if you are moving files from an IBM mainframe or a VAX, that you read the excellent paper, An Overview of Transporting SAS files between Hosts, on the SAS Institute web site at: <http://www.sas.com/techsup/download/technote/ts271.html>. Note that you should **not** use PROC CPORT to write files that are to be read by The system. This procedure creates files in an entirely different and incompatible format.

Reading SAS Transport Files

More than one data set may be stored in a single transport file. If The system finds more than one data set in a file, it will allow you to select the one that you want.

Writing SAS Transport Files

When the system writes a SAS transport file, it uses the file name of the input file, without the extension, as the internal name for the data set in the output file. The system will write only one data set to each output file.

Missing Data

SAS supports missing values. On input, all of SAS's missing value are converted to a single internal missing value in the system. On output to SAS, missing values are set to '.', the SAS standard missing value.

Input and Output Variable Types

The target output variable types selected by The system for each input variable type in a SAS Transport file are shown below in the first table. The second table shows the actual SAS Transport output variable type that results from each The system target variable type assigned during a transfer from some other format.

SAS Transport File Input to the System

Input Type	Default Target Type*
Numeric	Continuous unless print format is: DATE DDMMYY MMDDYY MMYY WORD, or WEEK, in which case it becomes a Character
Character Nominal	Character

*See Overriding Default Target Type Section

Output to SAS Transport Files from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

S-Plus Files

The system will read and write S-PLUS data sets. Files written on 64 bit machines such as DEC Alpha are not supported.

Standard Extension: [none]

Reading S-PLUS files

Because the S-PLUS file format is so unstructured that it allows the user to write almost anything, including code, into it, the system imposes a few restrictions on input files.

Specifically, your data should be in one of the following formats:

two dimensional matrix

S-PLUS list

dataframe

S-PLUS writes out its data in the native format of the machine on which it is running. This means that both the byte order and the width of numbers can vary between machines. On input, the system will automatically sense the byte order of the machine that wrote the file.

Writing S-PLUS files

On output, the system writes a S-PLUS dataframe. If your input data set does not have a variable named 'rownames', The system will create an extra variable containing the case number, stored as an integer variable and named 'rownames'. You can choose whether you want to write out a file with low to high byte order, appropriate for such processors as Intel or DEC, or a file with high to low byte order, for such processors as SPARC, HP, or Motorola. If you are using the Windows version of S-PLUS, select Intel (low to high) byte order on output.

Missing Data

S-PLUS supports missing values. On input, missing values are converted to the internal missing value symbol in the system. On output, missing values are converted to the value appropriate for each variable type.

Input and Output Variable Types

The target output variable types selected by The system for each input variable type in a S-PLUS file are shown below in the first table. The second table shows the actual S-PLUS output variable type that results from each system target variable type assigned during a transfer from some other format.

S-PLUS input to the System

Input Type	Default Target Type*
Integer Real Double	Continuous
Logical	Continuous
Character	Character

*See Overriding Default Target Type Section

Output to S-PLUS from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet

SPSS Data Files

The system will read and write SPSS data files from the following platforms:

Windows and OS/2

UNIX HP/Sun/IBM

UNIX DEC, and Standard Extension SAV

Reading SPSS Data Files

The system automatically recognises a file's platform of origin on input.

The system will read both compressed or uncompressed SPSS data files.

Writing SPSS Data Files

On output, the system allows you to choose a file type for Windows and OS2 or a UNIX file type either for the general group of HIGH-LOW (Sun, HP and IBM) or LOW-HIGH byte order machines (DEC).

The system always writes compressed files (which on typical survey data are notably smaller).

Value and variable labels are fully supported.

Missing Data

SPSS supports missing values. On input, all of SPSS's missing values are converted to a single internal missing value in the system. On output to SPSS, missing values are set to the SPSS system missing value.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a SPSS data file are shown below in the first table. The second table shows the actual SPSS data file output variable type that results from each system target variable type assigned during a transfer from some other format.

SPSS Data File Input to the System

Input Type	Default Target Type*
Number	Continuous
Number with date format	Character
Character	Character

*See Overriding Default Target Type Section

Output to SPSS Data Files from the System

Target Type	Output Type
Continuous	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Ordinal	
Nominal	
Integer	
Character	Character
Nominal	

SPSS Portable Files

SPSS Portable files (previously called Export files) were designed to transfer SPSS data sets between different kinds of computers. You can use them to move your data to and from mainframe SPSS-X and SPSS for the PC.

Standard Extension: POR

Reading SPSS Portable Files

Mainframe Portable files should be transferred to your PC using an error-correcting communications protocol. It is quite difficult to check these files visually for errors and certain errors may fatally affect the ability of the system to interpret the file.

Writing SPSS Portable Files

When the system writes Portable files, it does so with up to ten base thirty digits of precision.

Missing Data

SPSS supports missing values. On input, all of SPSS's missing values are converted to a single internal missing value in the system. On output to SPSS, missing values are set to the SPSS system missing value.

Input and Output Variable Types

The target output variable types selected by The system for each input variable type is in a SPSS Portable file are shown below in the first table. The second table shows the actual SPSS Portable output variable type that results from each The system target variable type assigned during a transfer from some other format.

SPSS Portable File Input to the System

Input Type	Default Target Type*
Number	Continuous
String	Character

*See Overriding Default Target Type Section

Output to SPSS Portable Files from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

STATA Files

The system will read and write data for any version of STATA including versions running on UNIX and the Macintosh.

Standard Extension: DTA

Reading Stata Files

The system can read data from any version of Stata. Character variables and dates are fully supported. Variable and value labels are transferred out of Stata.

Writing Stata Files

Stata holds the entire data set in memory. The system will therefore attempt to conserve as much space as is possible.

However, when The system is transferring from a format in which the width and number of decimal places are known (such as SPSS, dBASE, and worksheet formats), or when it is optimising the output variables, it will use the available information to minimise the size of your Stata data set. You can, of course, fine-tune this process by selecting types for output variables yourself.

Any variable and value labels present in the input data set will be written to Stata files.

Dates are written to Stata's internal date format.

Missing Data

Stata supports missing values.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a Stata file are shown below in the first table. The second table shows the actual Stata output variable type that results from each the system target variable type assigned during a transfer from some other format.

Stata Input to the System

Input Type	Default Target Type *
Byte Int Long Float Double	Continuous
Character	Character

*See Overriding Default Target Type Section

Output to Stata from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

Statistica

Standard Extension: STA

Reading Statistica Files

Statistica variable names can be up to 15 characters in length.

Version 5 can be read by the system.

Writing Statistica Files

Version 5 can be written by the system.

Any variable name in the source data set containing a left-parentheses followed by a number will be transferred into a STATISTICA subscripted variable.

Users should note that the STATISTICA error message, “You are trying to read an empty file”, will occur when STATISTICA cannot find a data file. Your STATISTICA files should be in the default drive or directory.

NOTE: Multiple worksheets stored in one file are not supported.

Missing Data

STATISTICA supports missing values.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a STATISTICA file are shown below in the first table. The second table shows the actual STATISTICA output variable type that results from each the system target variable type assigned during a transfer from some other format.

STATISTICA Input to the System

Input Type	Default Target Type*
Numbers	Continuous
Character	Character

Output to STATISTICA from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

Systat Files

Standard Extension: SYS

Reading Systat Files

When the system reads SYSTAT data sets, it processes the variable names by 1) dropping the dollar signs on character variables and 2) removing the parentheses before and after subscripts. For example, SCALE(1) becomes SCALE1.

Writing SYSTAT Files

Any variable name in the source data set containing a left-parentheses followed by a number will be transferred into a SYSTAT subscripted variable.

Users should note that the SYSTAT error message, "You are trying to read an empty file", will occur when SYSTAT cannot find a data file. Your SYSTAT files should be in the default drive or directory.

Missing Data

SYSTAT supports missing values.

Input and Output Variable Types

The target output variable types selected by the system for each input variable type in a SYSTAT file are shown below in the first table. The second table shows the actual SYSTAT output variable type that results from each the system target variable type assigned during a transfer from some other format.

SYSTAT Input to the System

Input Type	Default Target Type*
Numbers	Continuous
Character	Character

*See Overriding Default Target Type Section

Output to SYSTAT from the System

Target Type	Output Type
Continuous Ordinal Nominal Integer	Numeric format is determined by the settings in the Variable Attributes dialog box in the datasheet
Character Nominal	Character

Index

- (X+a) **p function, 19
- abs function, 18
- absolute deviation function, 24
- Adjusted Correlation Squared, 83
- AIC, 76
- Akaike Information Criterion, 76
- Alignment, 8
- Analysis – Frequency Table, 157
- Analysis of Variance, 164
- ANOVA, 111
 - Box-Cox Diagnostic Plot, 124
 - Formulae, 118
 - Scatterplot of Group SDs vs Means, 123
- ANOVA output
 - Brown-Forsythe Test, 120
 - Confidence Intervals, 119
 - Descriptive Statistics, 119
 - Kruskal - Wallis Non-parametric Test, 120
 - Levene's Test for Equal Variances, 119
 - Welch Test, 121
- ANOVA table, 59, 70
 - one-way ANOVA, 118
- Append variables, 8
- Appendices, 187
- arccos function, 20
- arcsin function, 20
- arctan function, 20
- Automatic Stepping Dialog Box, 69
- Bar Chart, 131
- Basic Attributes, 7
- Bayes Information Criterion, 76
- BIC, 76
- Box Plot, 129
- BoxCox function, 20
- Case Frequency, 10
- Case Label, 10
- Categorizing Continuous Variables, 154
- Change Model Dialog Box, 67
- Change Model Parameters, 170
- Change Pool, 170
- Checking for Independence, 147
- Checking for Normality with the Histogram, 145
- Checking for Normality with the Normal Probability Plot, 146
- Chi-square test, 93
 - function, 30
- Coefficient of Variation, 45
- Combining Categories, 152
- Common Transformation definitions, 18
- Components of Chi-square, 90
- Components of Likelihood Ratio G-square, 90
- Confidence Interval for the mean, 44
- Confidence Limits, 101, 104, 107
- Continuous variable, 8
- Contrast Options, 116
- Contrasts in ANOVA, 122
- Cook's Distance, 82
 - vs Predicted Value, 80
- Copy Attributes, 7, 11
- Corr/Cov Matrices, 71
- Correlation, 56
- Correlation Matrix
 - of Coefficients, 72
 - of Variables, 71
- cos function, 20
- Covariance Matrix
 - of Coefficients, 72
 - of Variables, 72
- Cramer's V, 92
- Creating a Bar Chart, 131
- Creating a Boxplot, 130
- Creating a Histogram, 128
- Creating a Means Comparison Chart, 133
- Creating a Normal Probability Plot, 136
- Creating a Scatterplot, 126
- Cumulative Distribution function definitions, 29
- Custom Plots, 73
- Custom transformations, 150
- Cutpoints, 160
- Data Management - Introduction, 3
- Data screening – missing values, 137

- Decimal Places, 8
- Define Variables, 35
- Defining Design Variables, 34
- Defining Variables, 31
- Delete Vars, 34
- Deleted Studentized Residual, 81
- Descriptive Regression Model, 51
- Descriptive Statistics, 41, 153
- Design Variables, 34
 - defining, 34
 - in multiple regression, 65
- Detrended Normal Probability Plot, 78
- deviations function, 24
- Diagnostic Plots, 72
- Diagnostics for Multiple Regression, 76
- Double Precision, 8
- Error Messages, 15
- $\exp(e^{*}X)$ function, 18
- F function, 30
- F test, 59
- Field Width, 8
- Fisher's Exact Test, 94
- Fixed effect model, 170
- Format, 8
- Frequency Analysis, 87
 - Differences, 90
 - Expected values, 90
 - Tables, 89
 - Tests, 90
 - Output Options, 89
- Frequency distribution, 28
- Frequency Distribution function definitions, 31
- F-to-Enter, 84
- F-to-Remove, 85
- F-value, 76
- Grouping variables, 13
 - cutpoints, 13
- G-square, 90
- Hat Diagonal, 81
 - vs. Predicted Value, 80
- Histogram, 127
- How to find Outliers, 140
- Hypotheses for two-way ANOVA, 113
- Hypothesis for one-way ANOVA, 112
- Import a File and Reading/Saving Data, 3
- Import/Export, 5
- Independence, 78
- Inference Statistics, 74
- int function, 19
- Integer variable, 8
- Intercept, 57, 68
- Interval variable, 10
- inv (1/X) function, 19
- Inverse Chi-square function, 30
- Inverse F function, 31
- Inverse Normal function, 30
- Inverse Student- t function, 30
- Kappa statistic, 92
- Kurtosis, 47
- KW, 120
- lag function, 25
- Levene's test for equal variance, 102
- Likelihood Ratio Chi-square, 93
- Linearity, 79
- Link Manager, 35
- ln function, 18
- log function, 18
- Longitudinal Variables - Defining, 31
- Mahalanobis' Distance, 82
 - vs Predicted Value, 80
- Mallows' Criterion, 76
- McNemar's Test of Symmetry, 94
- Mean, 22, 43
- Means Comparison Chart, 133
- Median, 48
- median function, 22
- Method, 69
- min/max function, 22
- Minimum/maximum Z-Scores, 46
- Miss, 43
- Missing Value Symbol, 5
- mod (xmodulus(p)) function, 19
- Model Menu, 67
- Modify Categories, 7
- Multipass Transformations, 23
- Multiple Correlation, 83
- Multiple Correlation Squared, 83
- Multiple Regression, 61, 168
 - ANOVA table, 70
 - Creating Dummy Variables, 64
 - Entering variables in, 61
 - Output Options, 66
- N , 43, 55
- N function, 22
- New Variables, 8
- NMissing function, 22
- Nominal variable, 8
- Normal function, 29, 31
- Normal Probability Plot, 135
 - of Residuals, 78
- Normality, 77, 145
 - Group, 47
- Normally Distributed Data, 146

- Notation, 42
- Number Missing, 43
- Number of rows per case, 6
- Odds Ratio, 91
- One Way ANOVA – Output Options, 118
- One-group *t*-Test, 163
- One-group *t*-tests, 98
- One-way ANOVA, 164
 - Model, 111
 - table, 122
 - Variances, 116
- Ordinal variable, 8
- Outliers, 79
 - finding and removing, 139
- p*_value, 59
- Paired *t*-test, 98, 162
- Parameter Confidence Intervals, 74
- Parameter Values, 57
- Partial Correlation, 83
- Partial Plots, 72
- Partial Regression Plots, 73
- Partial Residual Plots, 73
- Pearson *r*, 56
- Phi coefficient, 91
- Plots, 125, 171
 - for finding Outliers, 79
 - for One and Two-way ANOVA, 123
 - for Regression, 72
 - for *t*-test, 109
 - Boxplot, 109
 - Histogram, 109
 - Means Comparison Chart, 109
 - Scatterplot, 109
- population mean, 44
- Predicted Value, 81
- Predictive Regression Model, 51
- Quartiles and Interquartile Range, 48
- R*_sq, 56, 61
- Range, 45
- Rank Sum, 103
- Ratio
 - variable, 10
- Reference Group, 34
- RegMS, 59
- Regression, 51
 - Additional Output, 58
 - Analysis, 166
 - Coefficients, 84
 - Diagnostics, 73
 - Equation standardized, 58
 - Formulae, 55, 81
 - Model Interpreting, 167
 - Notation, 52
 - Options Menu, 74
 - Output - Descriptive Statistics Dialog Box, 70
 - Output - Simple Regression Scatterplot, 54
 - Statistics, 75
 - Variable Selection Criterion, 75
 - standardized, 58
- Regroup, 34
- Reorder Categories, 14
- Residual, 81
- Residual Mean Square, 56Residual vs Predicted Value, 79
- Residual vs. case #, 78
- Residuals vs Predicted Values, 80
- ResMS, 59
- Robust SD, 119
- Robust Statistics, 48
- Role, 8
- Rows separated by - Import/Export, 5
- R-square, 56
- Sample mean, 44
- Sample Size, 43
- Scatterplot, 125
- Scientific Notation, 8
- Screening and Changing Data in the System, 137
- SE Mean, 44
- Separate Variance *t*-test, 102
- Separating Columns character, 5
- Serial Correlation, 46
- Set Cutpoints, 7
- Shapiro and Wilk statistic, 48
- Show Group Names, 8
- sign function, 19
- Sign Test (Matched), 106
- Sign Test for One Group, 108
- Signed Rank (Matched), 107
- Signed Rank Test, 109
- Simple Linear Regression, 53, 166
 - Output Options, 55
- sin function, 20
- Single Group *t*-tests Output Options - Descriptive Statistics, 107
- Skewed Data, 146
 - using median, 48
- Skewness, 47
 - standard error of, 47
- Slope, 57
- Spearman rank correlation, 105
- Specify Basic Attributes, 7

- Specify Contrasts in one-way ANOVA, 115
- Specifying variable attributes, 7
- sq (X^{**2}) function, 19
- sqrt function, 18
- Sqrt_ResMS, 56
- Square Root of Residual Mean Square, 84
- Standard Deviation, 44
- Standard Diagnostic Plots, 77
- Standard Error, 57
 - of Mean, 44
 - of the Coefficients, 84
- Standardized Deviates, 90
- Standardized Regression Coefficients, 74
- Standardized Regression Equation, 58
- Statistical Data Analysis, 153
- Statistics for Normality checking, 146
- StdDev function, 22
- Step by step Intervention, 69
- Studentized Residual (deleted), 81
 - vs. Predicted Value, 80
- Student- t function, 30
- sum function, 22
- Summary function definitions, 22
- Summary Transformations, 21
- Supported file types, 4
- Surround character variable character, 5
- t- and Non-parametric Tests, 97
- Tables, 87
- Tabulated data, 87
- tan function, 20
- Test of Normality, 48
- Tolerance, 68
- Transformations
 - Multipass Functions, 23
 - Cumulative Distributions, 26
- Transforming Data for an Analysis, 149
- Transforming Variables, 15
 - Common and Trigonometric Transformations, 16
- Transforming Variables - Operators, 15
- Trigonometric, 20
- Trim_Mean, 49
- Trimmed ANOVA, 119
- trimmed function, 25
- Trimmed Mean, 49
- Trimmed Means Test, 104, 108
- Trimmed t -test, 105
- t -tests and Non-parametric Descriptive Statistics, 101, 104
 - for One Group, 107
 - for Paired Groups, 104
- Output Options, 101
- Single Group Output and Options, 107
 - statistics for One-group t -test, 100
 - statistics for Paired t -test, 100
 - statistics for two-group t -tests, 99
- Tests for Two Independent Groups, 102
- Tutorial, 137
 - Outliers, 138
 - t- and Non-parametric Tests, 158
- t-value for Coefficients, 75
- Two-group Output Options, 101
- Two-group t -test, 97, 159
- Two-way ANOVA, 165
 - Variances, 117
 - Model, 112
 - Output, 121
- two-way tables, 87
- Types of specified contrasts, 115
- Understanding the Link Manager, 179
- Uniform function, 31
- univariate summary statistics, 41
- User-defined, 15
- Using Frequency Analysis, 88
- Using One-way ANOVA, 114
- Using Simple Linear Regression, 53
- Using t- and Non-parametric Tests, 97
- Using Two - way ANOVA, 116
- Variable - copy attributes, 11
- Variable - modifying categories, 14
- Variable - setting cutpoints, 13
- Variable role, 10
- Variable Selection Criteria for Regression, 80
- Variable Type, 8
- Variable X model, 170
- Variables - Define Design Variables, 34
- Variables List Box, 34
- Variance, 45
- W_Stat, 48
- Warn if [] % of cases excluded, 69
- weights, 44
- winsorized function, 25
- Winsorized Standard Deviation, 49
- Yates' Corrected Chi-square, 94
- z-scores function, 25