

Minitab Basics

*for students in Biometry
(ZOO 631)*

About this guide

This handout summarizes the procedures to be used during the course. It does not cover every possible way of doing a particular procedure, nor does it provide every detail about these procedures; use the on-line help if you want to learn more.

This guide was written for Release 14; much of it applies to earlier releases but two very useful features — the project manager and editable graphs — are not available in older releases.

The material is presented in the same order as in the course, except that:

- Power analysis and other methods relevant to designing studies are presented in the last section, along with miscellaneous methods relating to probability distributions; in the course these are encountered at various points throughout.
- Bar charts for displaying categorical variables are presented in the “*Describing distributions*” and “**Describing relationships**” sections but in the course these may be deferred to the final section.
- Although the specific resampling macros are covered at the appropriate points in the various sections of this guide, general procedures for using macros in Minitab, and how to obtain these resampling macros, are described in a brief Appendix.

USING MINITAB

Graphical user interface

All, or nearly all, of Minitab's capabilities — certainly everything used in Biometry and Advanced Biometry, with the important exception of the **resampling macros** — can be accessed through its menu and dialog-window “graphical user interface.” In a standard installation of Minitab this indeed is the only interface available with the default settings. This guide is written entirely with reference to this interface, except for the descriptions of the resampling macros; these must be run using the command-line interface in the Session Window.

This interface is a very standard Windows interface, so in this guide I assume you know how to use it, with the exception of the following tip:

Dialog defaults

Any entries made in a dialog window — selection of variables, specification of options, etc. — remain in effect for later uses of that dialog window until you change or remove them. This includes any entries or selections made on a secondary dialog, e.g. that opened by the Graphs menu from the main dialog window of many statistical procedures: these remain in effect for later invocations of that procedure even if you do not re-open that secondary dialog.

This can be very convenient, eliminating the need to repeat choices when repeating a procedure, perhaps with some modifications; it can also be very annoying if you no longer want but forget to cancel the selections still in effect from previous uses of the procedure.

Command-line interface

Commands can be entered in the Session Window; for commonly used procedures you may find it easier and faster to do this than to work through several layers of menus and fill in multiple entries in dialog windows.

Enabling commands

To use commands the command-line interface must be enabled. If it is enabled a MTB> prompt will be present in the window. If it is not, to enable it:

1. make the Session window the active window (click in it)
2. open the main Editor menu by clicking on it

3. click on Enable Commands.

Commands are typed at the prompt, and executed by pressing the Enter key.

Abbreviations

Only the first four letters of any Minitab keyword (command, option, etc.) are used. This means that you can do any of the following:

- type only the first four letters
- type the full commands correctly
- type anything you want after the first four letters, including intervening words to make the command seem more like English, e.g. 'regress Y on 1 dependent variable, X, and store residuals in column C3' rather than 'regr Y 1 X C3'.

Subcommands

Some commands can or must be followed by subcommands. To tell Minitab that a subcommand is going to follow, end the main command with a semi-colon ';'. Rather than executing the command immediately, Minitab will respond with a SUBC> prompt. If additional subcommands are to follow, end this subcommand with a semi-colon also. The last subcommand must end with a period.

Learning commands

One easy way to learn the commands is simply to enable commands and then use the menu-window interface: Minitab will enter the corresponding commands in the Session window where you can study them. When it does this Minitab typically includes subcommands which are not necessary, so you can try simpler ways of invoking the procedures.

DATA ORGANIZATION, MANIPULATION, AND IMPORT/EXPORT

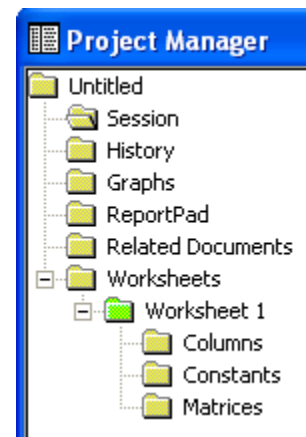
Projects

Minitab organizes your work into “projects,” which can be saved and opened. Project files are identified by the filename extension .MPJ.



The “Project Manager” window (right), opened by clicking the icon to the left, can be used to work with the various parts of a project. Projects include

- one or more worksheets of data,
- graphs (if any have been made),
- the “Session” window which contains a log of everything you have done in that project, including all statistical results,
- the “History” window, which contains a log only of the commands you have invoked, and
- the “ReportPad,” a very nice tool for collecting and presenting the useful output of a project. (see “Exporting results” on page 12)



Worksheets

Minitab shows the data in one or more worksheets much like spreadsheets in a program such as MS Excel. Worksheets ordinarily are saved and opened automatically as parts of projects, but if desired they can be saved or opened individually. The default filename extension is .MPW. There also is a special “Minitab portable” format for worksheet storage, indicated by the filename extension .MTP, which can be opened by other versions of Minitab (older releases or on other kinds of computer); the Minitab versions of the data sets on the textbook’s CD (and website) are in portable format.

Columns have names and types. Naming is quite flexible, with spaces and some special characters allowed. In dialog windows, columns can be referred to by name (in single quotes if the name includes spaces or special characters) or by column number (as e.g. ‘C1’). The data types are text, numeric, and date/time; text and data/time columns are indicated by ‘-T’ or ‘-D’ after the column number.

In Minitab worksheets, *rows are not treated as observations which should be kept intact*. In other words, different cells in a row may not have anything to do with each other, and so can be moved, sorted, deleted, etc., independently of each other. If in fact some or all of the cells in a row represent different variables for a single observation, and therefore should be kept together, *it is up to you to be sure any manipulations of the worksheet maintain the proper relationships between cells*.

The columns of the worksheet are the primary data entity, but there are two other types of data: constants (single numbers), and matrices (2-dimensional arrays). Minitab does not show these in the worksheet, but does list them in the Project Manager, and can display their values.

Entering data

Data can be entered directly into the worksheet from the keyboard or by cut-and-paste from other applications.

Data saved from various spreadsheet and database programs can be imported using the menu sequence

File ⇒ Open Worksheet...

(Do not use Open Project, which is only for Minitab projects.)

Text files

If data are in a simple text (“ASCII”) file with columns separated by tabs, Open Worksheet usually works. If, however, columns are separated by spaces, you need to use

File ⇒ Other Files ⇒ Import Special Text...

Exporting data

Data sets can be exported in a wide variety of formats, including text, older versions of Minitab, various versions of Excel, and other spreadsheet and database formats.

File ⇒ Save Current Worksheet As...

specify the File name: to export it to, and select the desired format from the Save as type: pull-down list.

Text files

Data also can be exported as a simple text (“ASCII”) file using

File ⇒ Other Files ⇒ Export Special Text...

This allows you to select which columns you want exported, and optionally to specify the format of the output file (using Fortran-style formatting syntax).

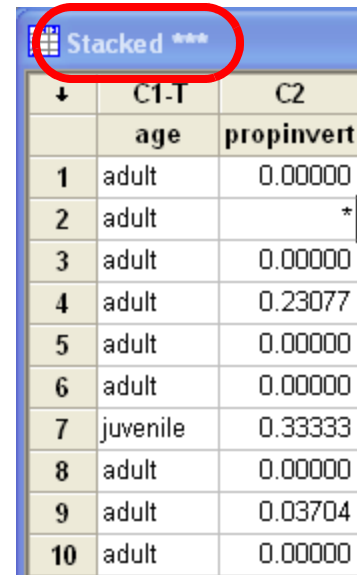
Data set manipulation

Stacked' and 'unstacked' data layouts

Before discussing the various ways data in worksheets can be manipulated, a basic issue of data organization needs to be explained. This pertains to how observations of a given variable but from different “groups” — experimental treatments, samples, or other categorizations — are arranged in the worksheet.

Stacked

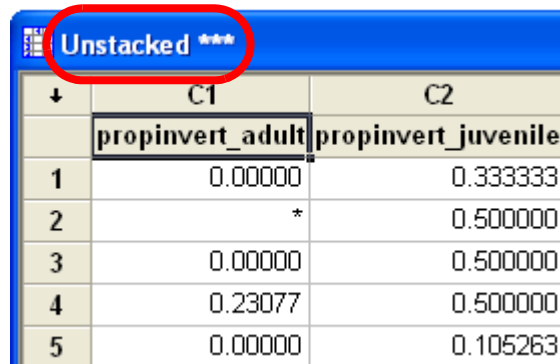
All the observations of a given variable might be in one column, with one or more other columns defining the different groups. For example, the data set shown (partly) to the right contains observations on a sample of fish. Each fish was classified as adult or juvenile, and the proportion of invertebrates in its stomach contents was recorded. In this worksheet all the proportions are “stacked” in one column (propinvert), with another column (age) identifying whether the value in that row was from an adult or a juvenile.



	C1-T	C2
	age	propinvert
1	adult	0.00000
2	adult	*
3	adult	0.00000
4	adult	0.23077
5	adult	0.00000
6	adult	0.00000
7	juvenile	0.33333
8	adult	0.00000
9	adult	0.03704
10	adult	0.00000

Unstacked

Alternatively, all the proportions for adults could be in one column (propinvert_adult), and all those for juveniles in another column (propinvert_juvenile), as shown here. If there were more than two categories of fish (e.g. adult-female, adult-male, juvenile), there would be a corresponding number of columns of values in the unstacked data.

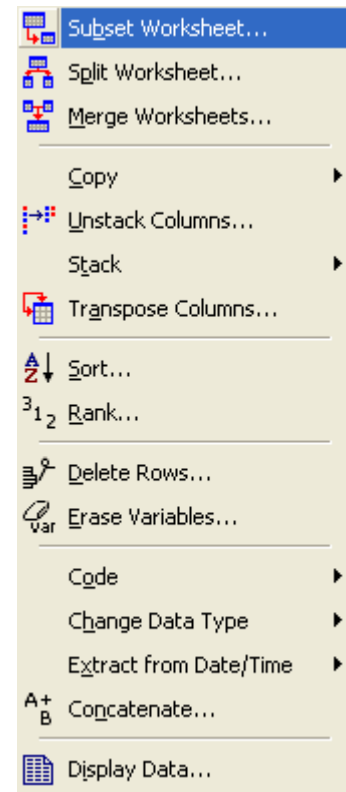


	C1	C2
	propinvert_adult	propinvert_juvenile
1	0.00000	0.333333
2	*	0.500000
3	0.00000	0.500000
4	0.23077	0.500000
5	0.00000	0.105263

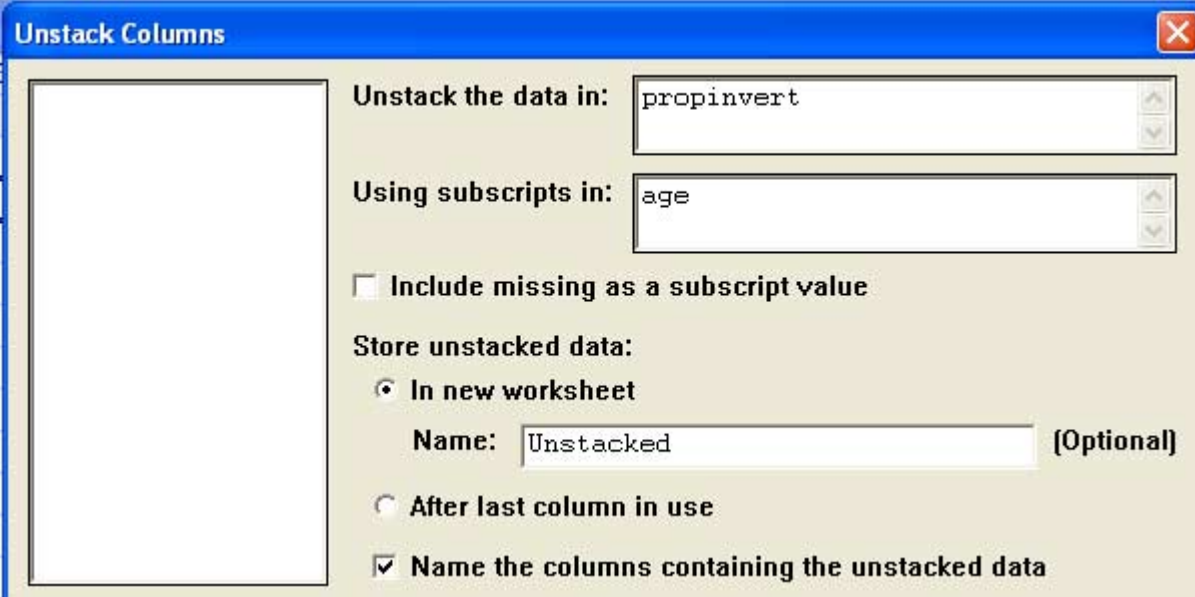
For most but not all purposes, the stacked layout is better; this is particularly true for analyses involving more than one “explanatory” (“independent”) variable, as dealt with in *Advanced Biometry*. In addition, for many procedures that can take data in either layout, how the procedures are specified depends on which layout the data are in. There are commands to convert data between these two layouts, described below.

Data menu

Most facilities for manipulating data are on the Data menu, with the Calculator on the Calc menu also performing data manipulations. The most useful of the items on the Data menu (shown at right) are:



- **Subset Worksheet...**
This creates a new worksheet containing only some of the rows of the current worksheet. The rows to be included (or to be excluded) can be specified by
 - listing their row numbers,
 - selecting the rows in the worksheet, using the mouse (“brushing” them), or
 - specifying a logical condition, in which case rows meeting the condition will be included (or excluded). Clicking the Condition... button opens a window resembling the Calculator (see below) in which the condition is created.
- **Split Worksheet...**
This creates two or more new worksheets, dividing the current worksheet according to one or more “By variables.” Each distinct value of the “By” variable (or distinct combination of values, if more than one “By” variable is used) defines a separate worksheet.
- **Copy ▶**
does the obvious, acting on columns, matrices, or constants.
- **Unstack Columns...**
splits a column (or a group of columns) into several new columns, according to the values of one or more other columns, i.e. converts from a “stacked” layout to an “unstacked” layout (see previous page).
The dialog for unstacking columns (see example on next page) asks which column(s) is/are to be unstacked and which column(s) define(s) the groups of observations to be separated. It also gives the choice of creating a new worksheet (which can be given a name) or of putting the new columns in the current worksheet, after the last column. The dialog shown below, applied to the “Stacked” worksheet on the previous page, produces the “Unstacked” worksheet at the bottom of the previous page.
- **Stack ▶**
This does the reverse of unstacking: it puts the contents of two or more columns into one column, one on top of the other. It can also create a second new column with values identifying which of the original columns a particular observation came from.
For example, the example unstacked data set shown above can be stacked to put all the proportions in one column, with a second column identifying whether



Unstack Columns

Unstack the data in:

Using subscripts in:

Include missing as a subscript value

Store unstacked data:

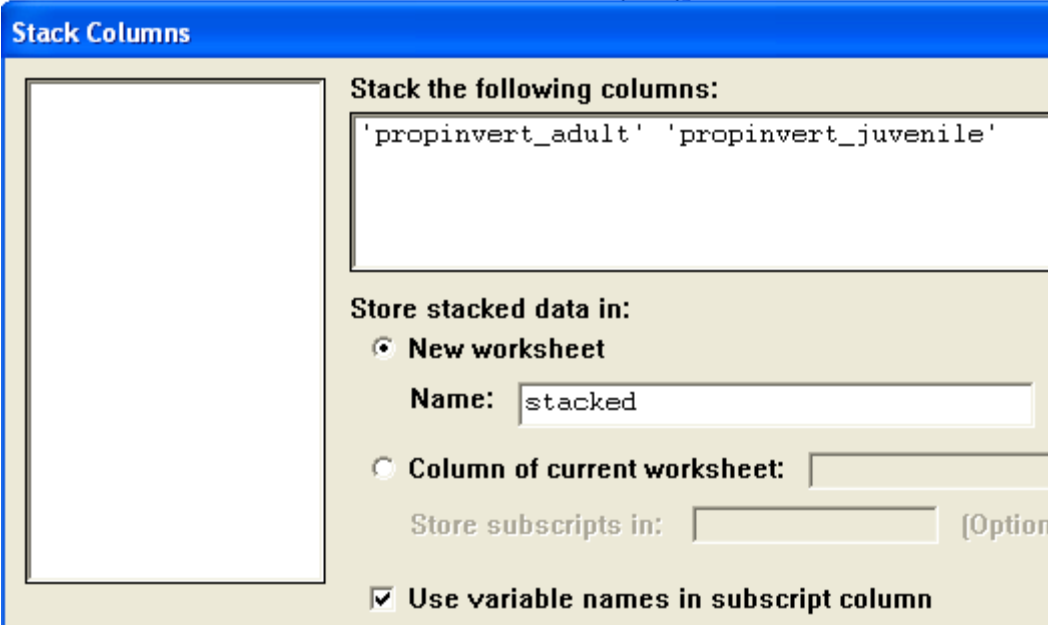
In new worksheet

Name: (Optional)

After last column in use

Name the columns containing the unstacked data

that row was from the adult column or juvenile column of the unstacked data set. The dialog (below) asks which columns are to be stacked, and whether to put them in a new worksheet or a new column of the current worksheet. For the example shown here, the worksheet produced by the stacking will be similar to the stacked worksheet shown on the previous page, except that all the adult values will precede the juvenile values (since entire columns are stacked).



Stack Columns

Stack the following columns:

Store stacked data in:

New worksheet

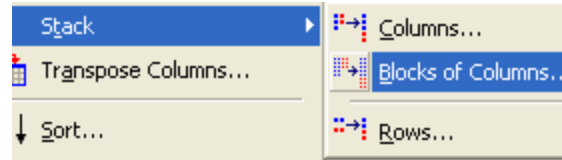
Name:

Column of current worksheet:

Store subscripts in: (Optional)

Use variable names in subscript column

It also is possible to stack blocks of columns, creating two or more new stacked columns. This option is selected from the small menu that pops up when **Stack** ▶ is selected, as shown here. The dialog for stacking a block of columns is considerably more complex than for creating a single stacked column, so I tend not to use it.

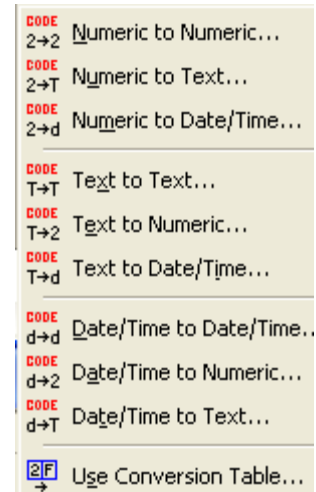


- **Sort...**
sorts one or more columns, in order of the values of one or more columns. If more than one column is sorted, cells in a given row are kept as a group. Note that this does not sort the entire dataset unless all the columns are specified, and therefore if done carelessly can destroy the organization of the data into observations.
- **Rank...**
creates a new column containing the ranks of the values — 1 for the smallest, 2 for the next smallest, etc. — in an existing column, without sorting or otherwise rearranging the data.
- **Delete Rows...**
does what it says.
- **Erase Variables...**
deletes columns, constants, or matrices.
- **Code ▶**

allows values of a column to be changed in a variety of ways. One or more values, or ranges of numeric or data/time values, are assigned a given new value, as in the dialog shown below. The new (coded) values can be put in the original column or (more safely) in a new column.

A particularly valuable use of Code is to replace cryptic numeric labels of samples or treatments with meaningful character labels (e.g. 'adult' and 'juvenile' rather than '1' and '2').

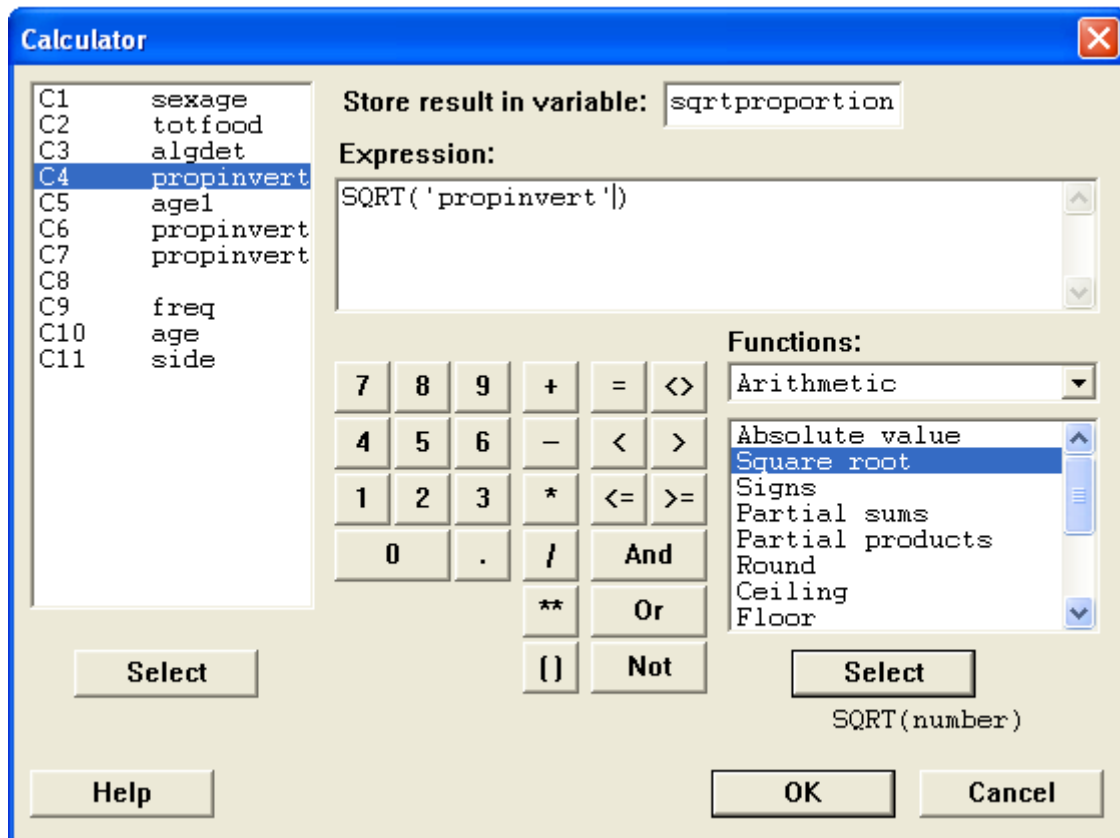
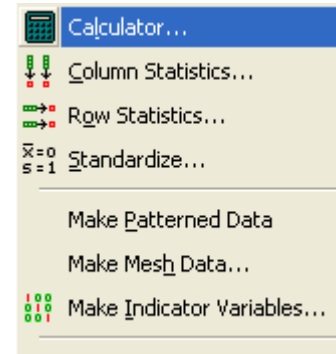
Code data from columns:	
sexage	
Into columns:	
age1	
Original values (eg. 1:4 12):	New:
1 2	adult
3	juvenile



Calculator

The Calc menu (part of which is shown to the right) contains one item, the Calculator, of great importance for data manipulation. Two other items on this menu, Random Data and Probability Distributions, as described in a later section. The other items are less useful and therefore not covered in this handout.

The Calculator window is used to transform or combine variables: to create a new column with values transformed from one or more existing columns. The column to be cre-



ated is specified in the Store result in variable: box, by giving either a column number (e.g. 'C12') or a column name (e.g. 'sqrtproportion'). The formula providing the values for the new variable is entered in the Expression: box. It can be typed in from the keyboard and/or using the number and operator buttons and scroll list of functions in the calculator window, and selecting existing variables from the list at the left. These expressions can be quite lengthy and complex, and the list of available functions is extensive, though of course often only a simple calculation is needed, as in the example above.

Exporting results

Text output from the Session Window, as well as graphs, can be transferred to a word processing program by cut-and-paste; for graphs, **Edit** ⇒ **Paste Special...** may work better than ordinary pasting.

Alternatively, if the Session Window is the active window,

File ⇒ **Save Session Window As...**

allows the entire Session Window to be saved in a variety of formats, including RTF and HTML. Similarly, if a graph is the active window

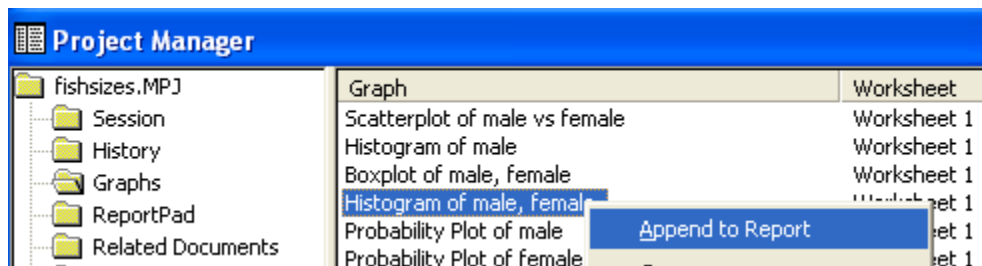
File ⇒ **Save Graph As...**

can export the graph in various formats, including JPEG and TIFF.

ReportPad

The “ReportPad” is essentially a WordPad document into which graphs or parts of the output in the Session Window can be copied. (WordPad is sort of stripped down version of MS Word, with moderate formatting facilities and the ability to save files in RTF format.)

The easiest way to add something to the ReportPad is to select the desired item (a graph or a section of the Session) in the Project Manager, then right-click on it and select Append to Report.



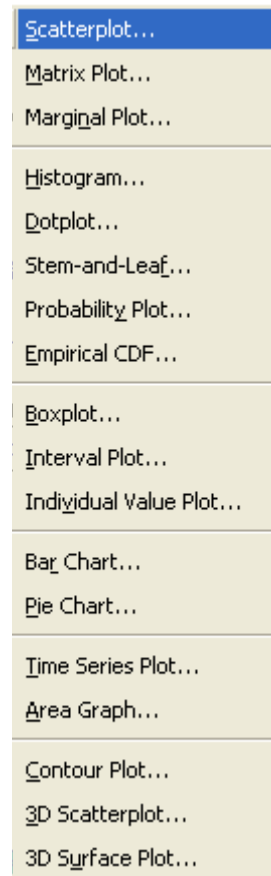
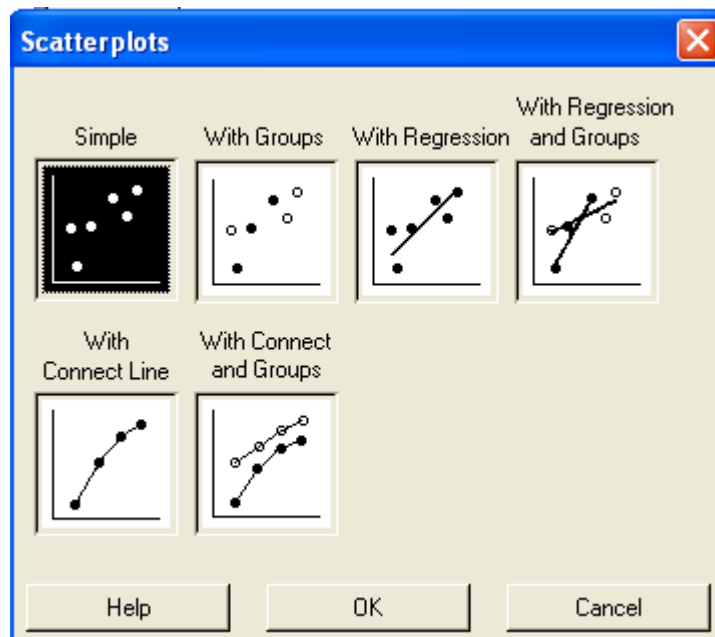
The report can then be edited as desired, including typing in additional text or pasting in items from other programs or files.

MINITAB GRAPHING: GENERAL FEATURES

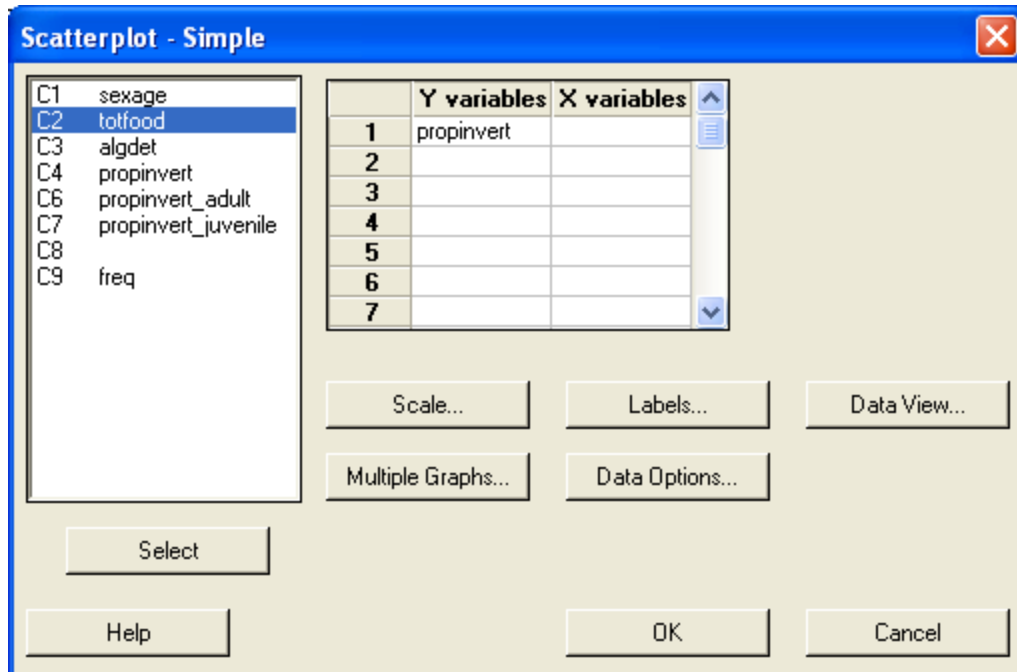
Minitab has very good graphing capabilities. They are not as advanced or extensive as in S-Plus or in specialized scientific graphing programs such as SigmaPlot. They are, however, entirely adequate for everything short of publication or presentation, and with some work on formatting can even be used for publication and presentation. In this section I describe some general aspects of creating and working with Minitab graphs; details for specific sorts of graphs are presented in the appropriate later chapters.

Creating graphs

Many of the statistical procedures on the Stat menu produce graphs, typically as option parts of their output, but the main tools for creating graphs are on the Graph menu, shown to the right. Most of the selections on this menu first open a diagrammatic window like the one below for scatterplots, in which you



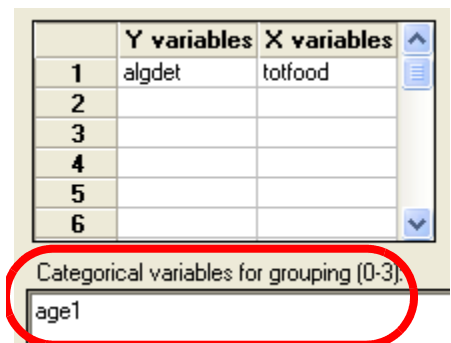
select the particular form of the graph you desire, by clicking on it and then on the OK button. This then opens a dialog window generally like the scatterplot dialog shown below.



Multiple graphs

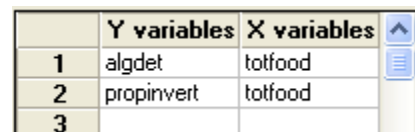
Multiple graphs can be created using any or all of four general methods:

- For most graph types the initial menu selection includes one or more “With Groups” options, as in the first scatterplot selection window on the preceding page. If one of these is selected, the graph dialog, in addition to needing the variables to be directly plotted, will require that one or more categorical variables be specified to define the “groups,” as shown here in a partial view of a “scatterplot with groups” dialog window.

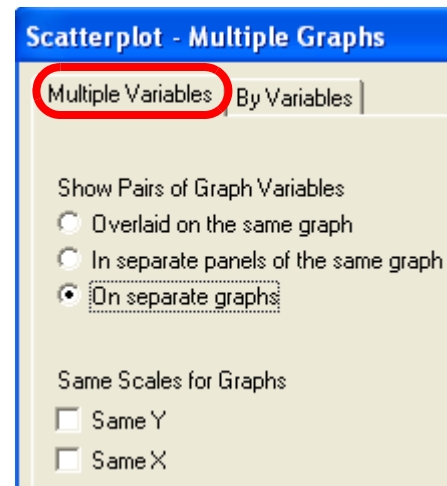


How these groups will be differentiated varies with the type of plot being made; for some, e.g. a scatterplot, different symbols are used, while for others, e.g. a boxplot, separate plots are made in the same graph axes.

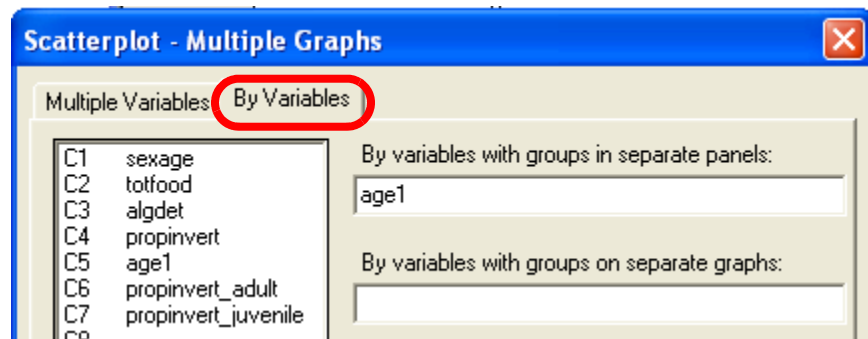
- Several different plots can be requested in the main dialog box, by listing multiple sets of variables as in the scatterplot example at the right.



Clicking the Multiple Graphs... button present on most graphing dialog windows then opens a window like the one to the right. On the Multiple Variables tab of this window you can choose whether to have each of the plots as a separate graph (i.e. in a separate window), as a panel in the same graph (i.e. in different parts of one window), or overlaid. If the plots are not to be overlaid, you can choose whether to have them all use the same scales for either or both axes; requiring this can facilitate comparison between graphs.



- Each requested plot can be divided into multiple graphs according to the distinct values of one or more “by variables.” To do this, click the Multiple Graphs... button on the main graph dialog window, then click on the By Variables tab. On this tab you can list one or more variables so that separate graphs will be created, either in separate panels of one window or in separate windows, for each different value (or combination of values) of the “by” variable(s). In the example at the top of the next page, each graph requested in the main dialog would have two panels, one for adults and one for juveniles (the two values of the age1 variable).

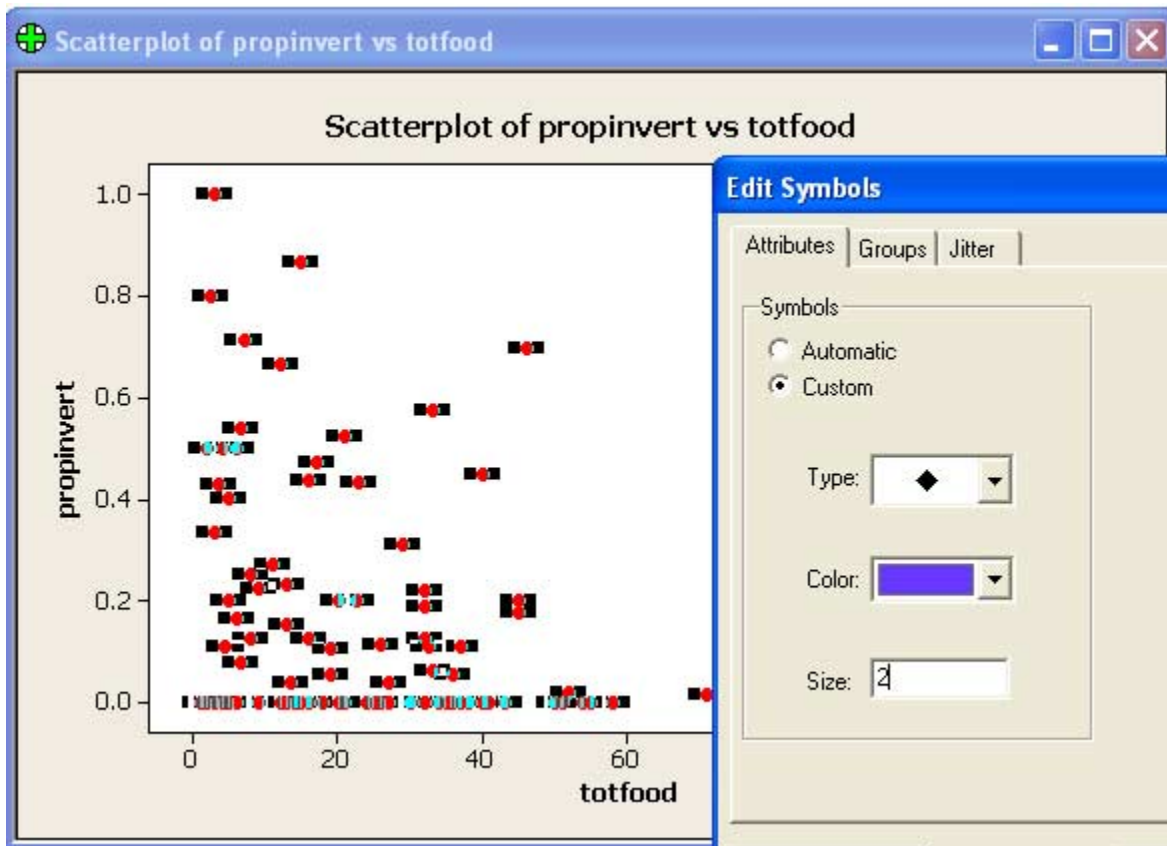


- After a graph is made, right clicking on it brings up the menu shown below under “**Editing graphs.**” One item on this is Panel..., which opens the “Edit Panels” window; the “Panels” tab of this window allows “by variable(s)” to be specified as in the method described in the preceding bullet.

Editing graphs

Selecting any part of a graph and then right clicking (or double-clicking) opens a dialog allowing that part to be modified. For instance, the text and/or the style (font, size, etc.) of an axis label can be changed, the scaling (range, tick placement, etc.) of an axis can be modified, or the symbol, size and color of all points in a plot — or of a single point — can be altered.

In the example shown here, the symbols used in a scatterplot are being changed in all three characteristics:

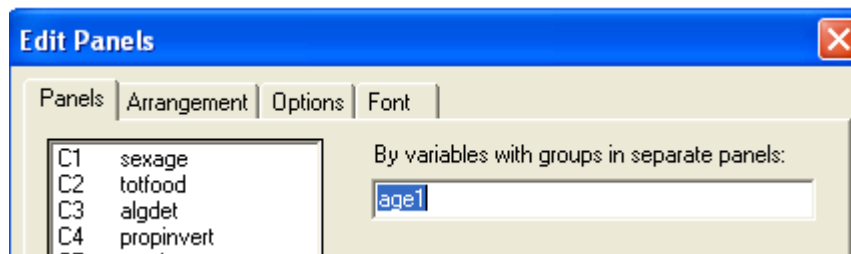
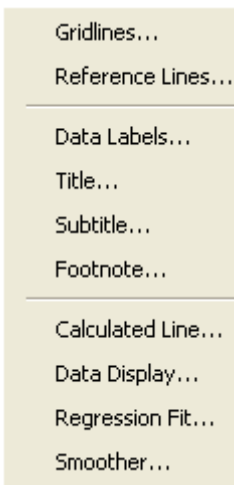
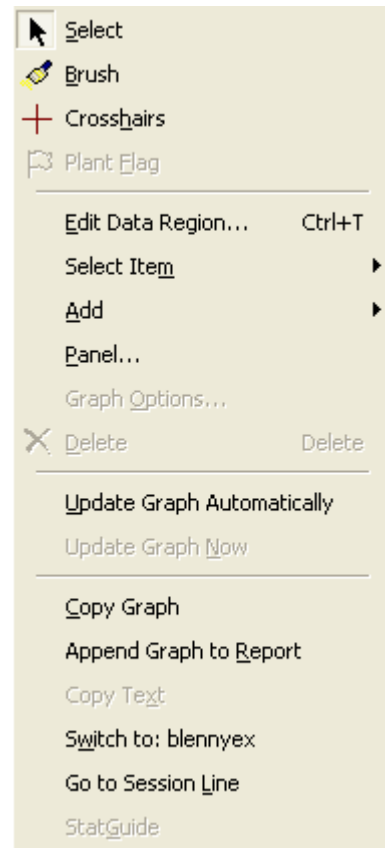


Graph editing menu

Right clicking in a graph, or selecting the main Editor menu when a graph is the active window, opens a menu like that shown on the next page. Which items on this menu will be available (not grayed out) will depend on what part of the graph was selected before right clicking; similarly the item shown here as Edit Data Region will edit whatever part of the graph was selected.

The most useful editing items on this menu are:

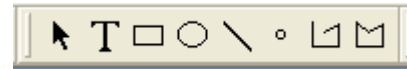
- **Edit xxx...**
opens all the tools for modifying whatever part of the graph is selected, e.g. the Edit Symbols window shown above.
- **Select Item ▶**
allows a different part of the graph to be selected (than had been selected before opening this menu).
- **Add ▶**
allows a variety of things to be added to the graph, as shown on the list to the left. What these will be will depend on what sort of graph is being edited. In all cases the middle group, having to do with annotation and labelling, is available. In most cases **Gridlines...** and **Reference Lines...** will be available; the latter are horizontal or vertical lines at user-specified values of the respective variable. **Data Display...** in the bottom group usually will be available; it opens a window allowing choices similar to those provided by the **Data View...** button when graphs are created. The other items in the bottom group here are specific to scatterplots.
- **Panel...**
splits a graph into several panels, one for each distinct value (or combination of values) of one or more specified “by” variables. The first tab of the Edit Panels window is where the “by” variable(s) is/are specified. Other tabs allow the arrangement of the panels (how many rows and columns of panels were window), and various aspects of their appearance, to be modified.



- Copy Graph
allows the graph then to be pasted into another application (use “Paste Special...”)
- Append Graph to Report
puts the graph in the ReportPad.

Annotation

A basic set of tools is available for annotating graphs, accessed from the Graph Annotation Tools toolbar (which may need to be activated on the Toolbars item on the main Tools menu). The tools are a text tool for adding text, and tools for adding various shapes; the colors, fill, thickness, etc. of these can be edited.

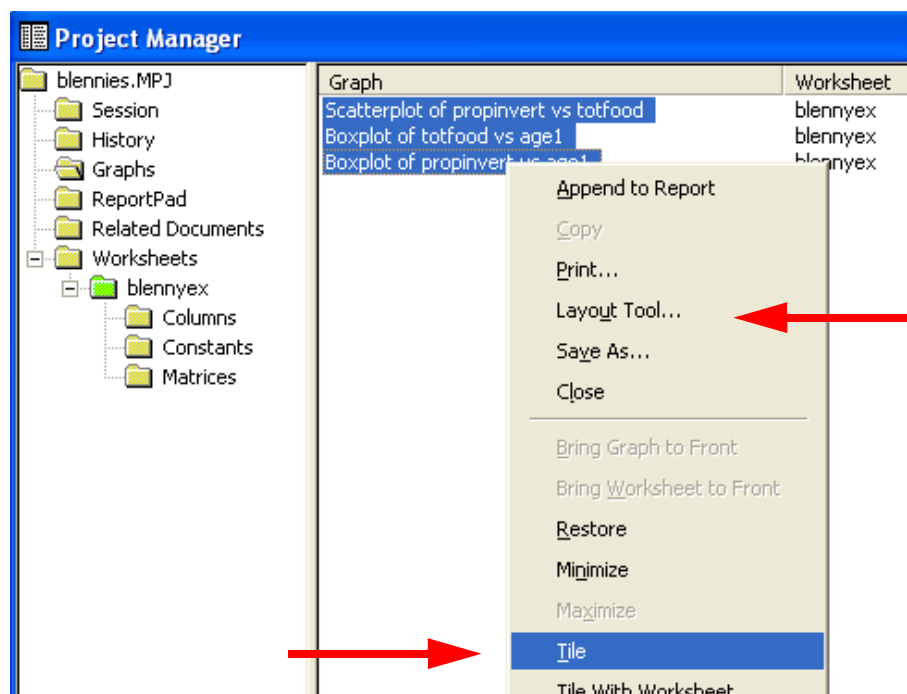


Layout

If one or more graphs are selected on the Graphs folder of the Project Manager, and then right clicked on, a menu opens with various actions that can be applied to the graphs, as shown at the top of the next page.

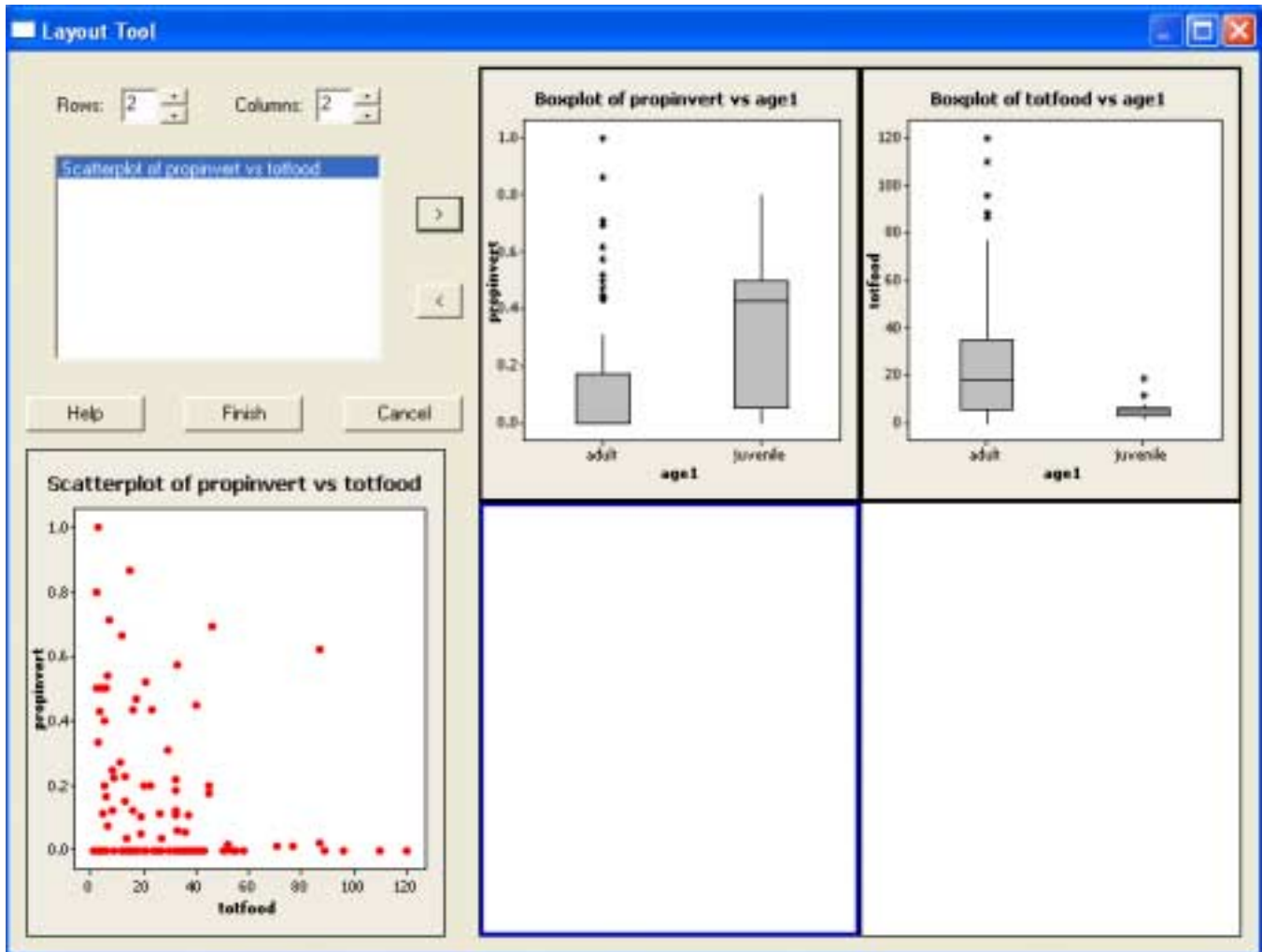
Tile

Two of these actions are “Tile” and “Tile with Worksheet.” These arrange the selected graphs in a grid filling the Minitab window; “Tile with Worksheet” puts the active worksheet in a panel of this grid of graphs.



Layout tool

The same menu gives access to the Layout Tool; it also can be accessed on the main Editor menu when a graph is the active window. This tool (shown on the lower part of the next page) allows multiple existing graphs to be put together as panels in one graph window. This composite graph then will be listed on the Graph folder of the Project Manager, allowing it to be exported, tiled, copied to the ReportPad, etc.

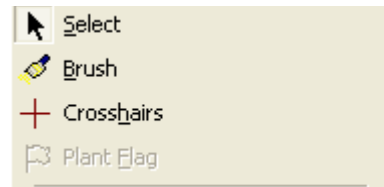


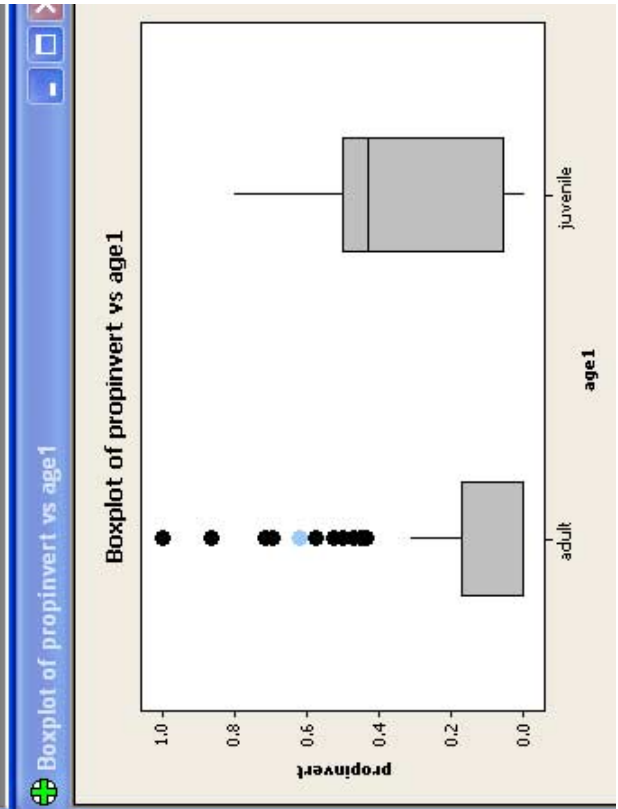
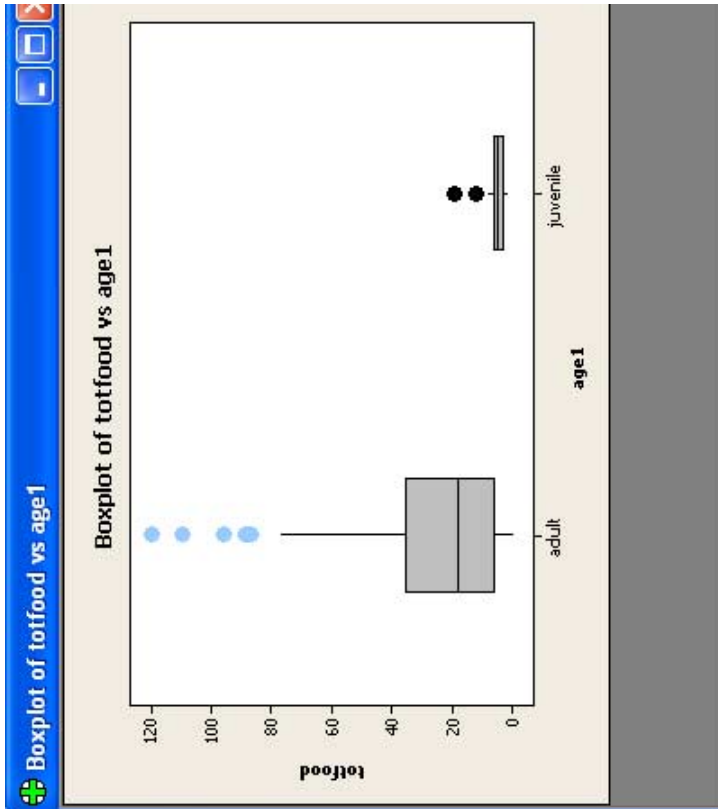
Interacting with graphs

An important part of modern data analysis is interactive exploration of graphs of the data, along with the data themselves. For instance, it is important to look for and assess the effects of “outliers.” A very useful tool for this sort of data exploration is graph “brushing.”

Brushing

The top part of the graph-editing menu discussed above (opened by right-clicking on a graph) includes a three- or four-way toggle determining the effect of the cursor. The second choice turns on interactive “Brush” mode. In this mode the cursor becomes a pointing finger, and any pointed selected in the graph will be shown in a different color. In addition, the row number of the selected point(s) will be listed in a little pop-up window and those rows will be flagged in the worksheet. And if any other graphs also are in “Brush” mode, the selected observation(s) will be highlighted in them as well. The example on the next page shows all of these features of brushing.



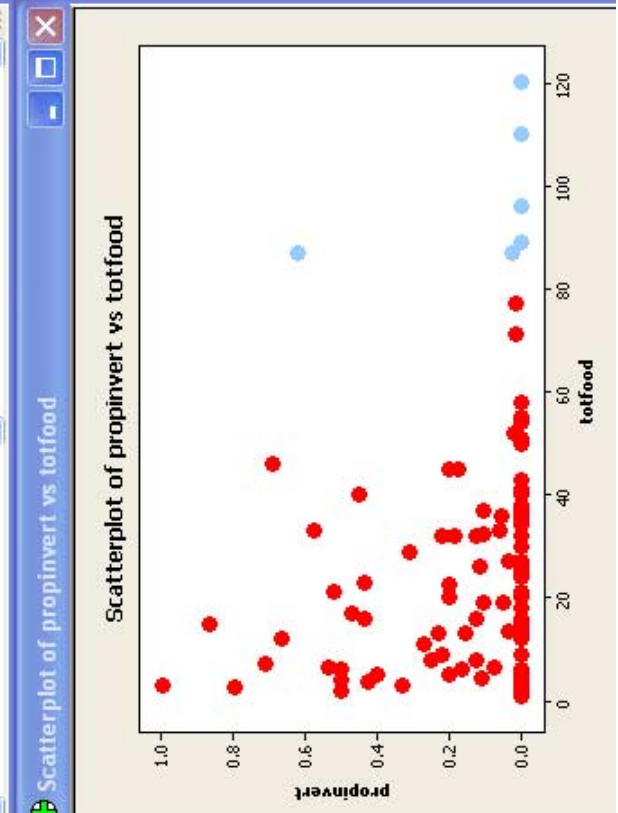


blennyex ***

	C6	C7	C10	age
	propinvert_adult	propinvert_juvenile		
7	0.00000	0.666667		
8	0.03704	0.500000		
9	0.00000	0.000000		
10	0.00000	0.000000		
11	0.00000	0.800000		
12	0.00000	0.500000		
13	0.00000	0.428571		
14	0.00000	0.250000		
15	*	0.000000		
16	0.00000	0.400000		
17	0.00000	0.000000		
18	0.00000			
19	0.00000			
20	0.05263		0	
21	1.00000		0	

Brushing

Row
13
14
90
105
116
121



DESCRIBING DISTRIBUTIONS

Plots of distributions

Histograms

To create histograms along with the descriptive statistics, use the preceding process

Stat ⇒ **Basic Statistics** ⇒ **Display Descriptive Statistics...**

then click on the Graphs... button, and in the next window select which plots are desired. These graphs can be modified as described below.

To create histograms without the descriptive statistics, use

Graph ⇒ **Histogram**

Click the box for a Simple graph, then specify the variable(s) for which the plots are desired.

To change the “binning” (number and location of the bars), open either the Edit Bars window (left click on the bars, then right click, and choose Edit bars...) or the Edit Scale window (left click one of the numbers on the X axis, then right click, and choose Edit X Scale.... In either of these windows, select the Binning tab and modify as desired. “Midpoints” are the middles of bars, while “Cutpoints” are the lower limits of bars. Their positions can be specified in the box at the bottom of the window either by entering all the desired values, or by the notation $a:b/c$, which gives a sequence from a to b in steps of c .

Comparing distributions

Overlaid

After Graph ⇒ Histogram..., click the box for a With Outline and Groups graph. If the distributions are in separate columns (unstacked), specify the columns in the Graph variables: box and check Graph variables form groups. If the distributions are stacked, specify the quantitative variable in the Graph variables: box and the categorical variables defining the groups in the Categorical variables for grouping (0-3): box; if doing more than one quantitative variable at once, uncheck Graph variables form groups unless you want all the histograms overlaid.

To make the groups easier to distinguish in these overlaid histograms, it helps to use thick lines for the outlines: left click on the bars, right click, select Edit Area..., select Custom under Borders and Fill Lines, then select a larger Size: (e.g. 4).

Separate

After Graph ⇒ Histogram..., click the box for a Simple graph, specify the quantitative variable(s) in the Graph variables: box, and click on the Multiple Graphs... button. If the distributions are unstacked, select the desired layout on the Multiple Variables tab; if the data are stacked, on the By Variables tab specify the grouping variable(s) in the By variables... box for the desired layout.

Boxplots

To create boxplots along with the descriptive statistics, use

Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics...

then click on the Graphs... button, and in the next window select which plots are desired.

To create boxplots without the descriptive statistics, use

Graph ⇒ Boxplot...

Click the box for a One Y; Simple graph, then specify the desired column(s) in the Graph variables box.

Comparing distributions

If the groups are in separate columns (i.e. the data are unstacked), after Graph ⇒ Boxplot..., choose Multiple Y's; Simple and specify the columns in the Graph variables box. If the data are stacked, after Graph ⇒ Boxplot..., choose One Y; With Groups, specify the quantitative variable(s) in the Graph variables box, and the column(s) defining the groups in the Categorical variables ... box.

Individual value plots

For small data sets the summarization implicit in histograms and boxplots may not be necessary: an “individual value plot” may be clear enough to use, without losing any details of the data.

Graph ⇒ Individual Value Plot...

Such a plot shows all the individual observations, with their values scaled along the vertical axis. As shown in the example on the next page, the points also are randomly scattered in the horizontal direction to reduce overlap; the degree of this “jittering” can be modified on a tab of the Edit Individual Symbols dialog opened by right-clicking on a point in the plot.

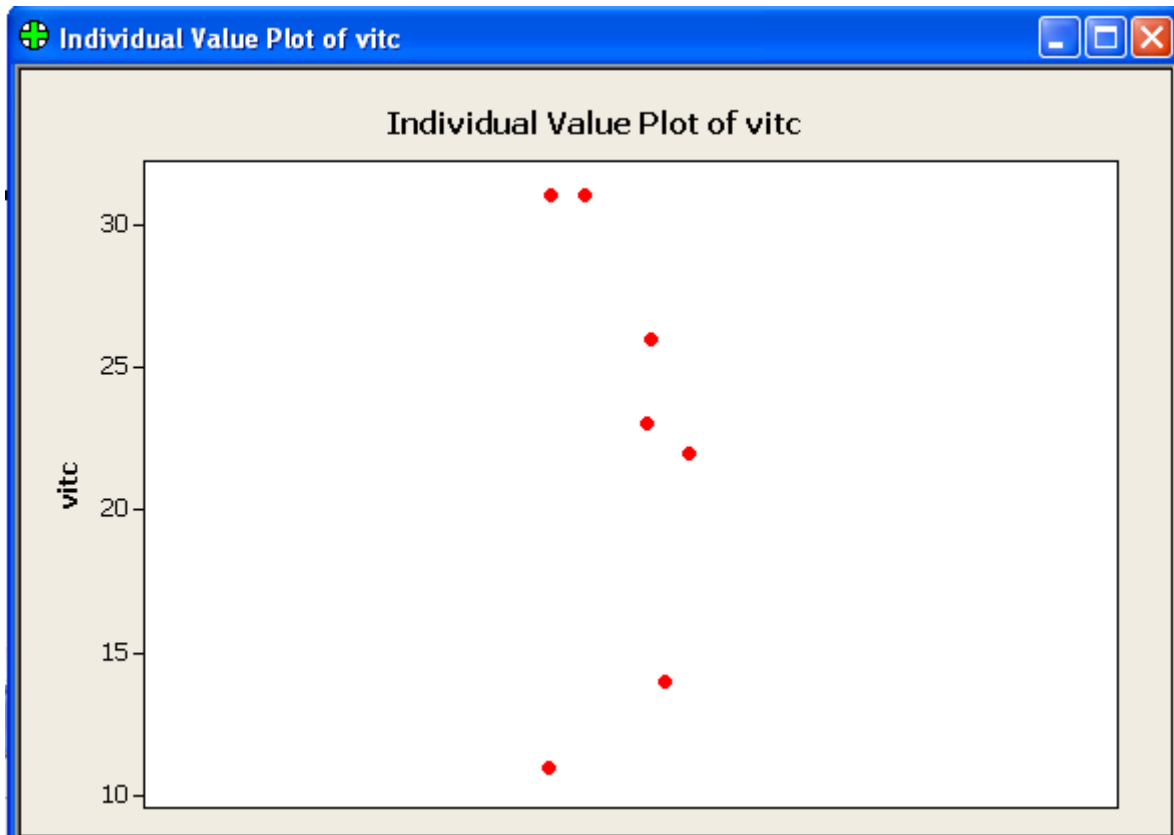
Normal quantile-quantile plots

Normal quantile-quantile plots can be made in various ways; the two easiest are as follows:

Stat ⇒ Basic Statistics ⇒ Normality Test...

and then select the variable, or

Graph ⇒ Probability Plot...



select a Single plot, and then specify the variable(s).

These two methods produce plots in which a straight regression line has been fit to the quantile-quantile points; the Probability Plot also shows confidence bands around this regression. I find the regression line and confidence bands distracting and so usually remove them after the plot is made, by selecting them on the plot and hitting the Delete key.

These two procedures, strictly speaking, produce probability plots rather than quantile-quantile plots. The only difference is in the scaling of the vertical axis; probability plots, rather than showing the values of the normal scores on a linear scale, instead show the percentile of each point, but on a non-linear probability scale that has the same effect as plotting normal scores.

These two methods also produce formal tests of the null hypothesis that the distribution is normal, along with summary statistics. **These tests rarely are useful, as will be explained in class.**

Comparing distributions

To get separate plots for different levels of a categorical variable (if the data are stacked), click on the Multiple Graphs... button, and on the By Variables tab specify the grouping variable(s) in the **By variables...** box for the desired layout.

Charts for categorical variables

The frequencies (or relative frequencies) of the different values of a categorical variable can be shown in bar or pie charts. Before describing how these charts are created, however, two ways such data can be organized in a worksheet need to be explained.

- Each observation may occur individually, as in the left example below.
- Frequencies of the categories may be given in a column, with each category as a row, as in the right example below.

barn_id	cause
1	smother
2	smother
...	
49	smother
50	undercut
...	
208	unknown

count	cause
49	smother
5	undercut
...	
38	unknown

Bar chart

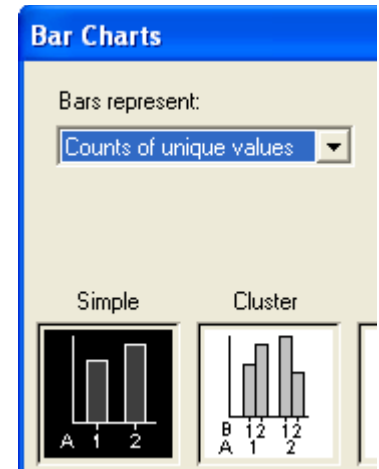
Graph ⇒ Bar Chart...

This opens a window in which two choices need to be made. For a single chart of a single categorical variable, select Simple chart. Then

- If the worksheet rows are individual observations, in the Bars represent: scroll box, select Counts of unique values as in the example here.
- If instead the worksheet contains frequencies, select Values from a table in this scroll box.

Then click the OK button.

The next window to open is for specifying the variable(s) to be charted.



- If the data are individual observations, the variable containing the values of the categorical variable is listed in the Categorical variables: box. (In the example data to the left above, this variable is “cause.”)
- If instead the data contain frequencies, the column of frequencies is listed in the Graph variables: box and the column with the category labels is listed in the Categorical variable: box.

The Bar Chart Options... button opens a window in which you can request that the vertical axis show percentages rather than absolute frequencies, and/or that the bars be ordered by their frequencies.

Pie chart

Graph ⇒ Pie Chart...

The choice of data layouts (as discussed above) is made by radio buttons at the top of the Pie Chart window.

- For individual values, select Chart raw data and then specify the variable in the Categorical variables: box.
- If the data set lists frequencies, select Chart values from a table, specify the Categorical variable: and then specify the column of frequencies in the Summary variables: box.

Descriptive statistics

A variety of statistics can be obtained individually, but the most convenient approach is to ask for a set of descriptive statistics as follows.

Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics

then specify the variables for which you want the statistics. (Plots also can be requested, as described above.)

The default output will be the mean, standard deviation, standard error of the mean, median, upper and lower quartiles, minimum, maximum, and numbers of non-missing and missing observations for each column specified. The selection of statistics can be changed by clicking on the Statistics... button.

Comparing distributions

To obtain descriptive statistics for subsets of the data, enter in the By variables box the variable(s) defining the groups (subsets) for which you want the statistics calculated.

DESCRIBING RELATIONSHIPS

This section focuses on relationships between **two quantitative variables**. Describing the relationship between **one quantitative** variable and one or more **categorical** variables amounts simply to comparing (between levels of the categorical variable) descriptions of the distribution of the quantitative variable, using methods presented in the “Comparing distributions” parts of the preceding section. Describing relationships between **two categorical variables** is covered at the end of this section.

Scatterplots

Graph ⇒ Scatterplot...

then select a Simple plot. Only the Y and X variables need to be specified. Specifying more than one Y–X pair will produce a separate plot for each pair.

LOWESS smoother

Graph ⇒ Scatterplot...

then select a Simple plot and specify the Y and X variables. Click the Data View... button, then the Smoother tab, and select Lowess. The smoothness of the LOWESS can be controlled by changing either or both of Degree of smoothing: or Number of steps;; increasing either makes the LOWESS smoother, while decreasing them makes it less smooth.

A LOWESS also can be added to an existing scatterplot by selecting (left clicking in) the plot, right clicking, and choosing Add ⇒ Smoother; the smoothness again can be modified if desired.

Distinguishing groups

It often is of interest to compare the relationship between two quantitative variables, among two or more groups of observations. (In effect, this examines the three-way relationship between the two quantitative variables and a categorical variable defining the groups.)

If the groups are unstacked, two methods are possible. The obvious one is to select, after Graph ⇒ Scatterplot..., a With Groups plot type, then enter the appropriate pairs of Y and X variables and select the X-Y pairs form groups checkbox. Alternatively, a Simple plot can be selected, the two or more pairs of Y-X variables specified, then click Multiple Graphs... button and select Overlaid on the same graph.

If the groups are stacked, there again are two methods. The obvious one is to select, after Graph ⇒ Scatterplot..., a With Groups plot type, enter the Y and X variables,

then enter one or more Categorical variables for grouping. Alternatively, grouping can be applied to an existing Simple scatterplot by selecting (right clicking on) the symbols, right clicking, selecting Edit Symbols..., selecting the Groups tab, then designating the appropriate Categorical variables for grouping.

LOWESSes (separate for each group) can be added to grouped scatterplots by either of the methods described above for a single LOWESS. If adding LOWESSes to an existing plot, check the Apply same groups ... checkbox on the Add Lowess Smoother window; if you forget and the LOWESS initially ignores the grouping, select the LOWESS (left-click on it on the graph), right click, select Edit Lowess Smoother..., click the Groups tab, and enter the appropriate Categorical variables for grouping.

Marginal plots

Graph ⇒ Marginal Plot...

produces a simple (ungrouped) scatterplot with either histograms, boxplots, or dot-plots of the individual variables along the margins of the scatterplot.

Correlation coefficient

Stat ⇒ Basic Statistics ⇒ Correlation

then specify the two (or more) variables

Regression

Stat ⇒ Regression ⇒ Regression...

then specify the Response and Predictor variables.

Regression plot

There are two ways to create a scatterplot with the fitted regression line.

Fitted line plot

This method gives only a scatter plot with a single fitted regression line; it cannot give separate fitted lines for different groups, and cannot be combined with such things as a LOWESS, or use of different symbols for different groups. It does, however, produce the regression output in the session window.

Stat ⇒ Regression ⇒ Fitted line plot...

then specify the Response and Predictor variables.

Various residual plots can be requested (Graphs button) and/or residuals and fits can be stored (Storage button) for later graphing.

Added to scatterplot

Similarly to the LOWESS smoother discussed above, a fitted regression line can be included in or added to a scatterplot. This method is more flexible than the preceding, since the scatter plot can be enhanced in any way desired; it does not, however, produce any regression output or even give the regression equation.

Graph ⇒ **Scatterplot...**

then select With Regression and specify the Y and X variables.

A fitted regression line also can be added to an existing scatterplot by selecting (left clicking in) the plot, right clicking, choosing Add ⇒ Regression Fit...

Residual plots

Using either Stat ⇒ Regression ⇒ Regression... or Stat ⇒ Regression ⇒ Fitted line plot..., residual plots can be produced immediately by clicking on Graph and selecting the desired plots. Alternatively, you can click on Storage, specify both residuals and fits to be stored, and then do separate plots as desired.

Categorical variables

Data layouts

There are at least three ways data for a two-way contingency table, i.e. describing the relationship between two categorical variables, can appear in the worksheet. What can be done with the data, as well as how it is done, depends on the layout.

Individual observations

Every independent observation can be in a separate row, with two (or more) columns containing the categorical variables by which each observation is cross-classified, as shown in the example to the left below.

Stacked frequencies

Each combination of levels of the categorical variables can be in a separate row, with columns (in the example below right, **cause** and **site**) containing the categorical variables defining these combinations, and an additional column (**count**) containing the frequencies of the category combinations.

individual observations

barn_id	cause	site
1	smother	above
2	smother	above
...		
49	smother	above
50	undercut	above
...		
208	unknown	below

stacked frequencies

count	cause	site
49	smother	above
5	undercut	above
...		
38	unknown	below

Unstacked frequencies

contingency-table layout				
site	smother	undercut	crowded	unknown
above	49	5	8	19
at	11	19	4	23
below	13	17	2	38

The row-and-column layout of a contingency table can be directly represented in the rows and columns of the worksheet, with separate columns containing different levels of one of the categorical variables and separate rows representing different levels of the other variable.

Stacked or grouped bar charts

Stacked or grouped bar charts can be made with data in either of the first two layouts above, as individual observations or stacked frequencies; to the best of my knowledge **these graphs cannot be made with data as unstacked frequencies.**

Individual observations

Use

Graph ⇒ Chart...

In the Graph Variables box, set Function to Count, Y to any quantitative variable present in every observation (in the example data set above, it could be “barn_id”), and X to the variable (*e.g.* “site”) for the levels of which you want separate bars in a stacked bar chart, or separate groups of bars in a grouped bar chart.

In the Data display box, set For each to Group and Group variables to the variable (*e.g.* “cause”) defining the categories within each bar, or the bars within a group of bars.

Stacked bar chart

To have the categories of the Group variable displayed as segments of the bars (and a bar for each category of the X variable), select Options, then for Groups within X select Stack and enter the variable defining the categories within each bar (*e.g.* “cause”), and select Total Y to 100% within each X category.

Grouped bar chart

To have the categories of the Group variable displayed as separate bars (and a separate cluster of such bars for each category of the X variable), select Options, then for Groups within X select Cluster and enter the variable defining the bars within each group of bars (*e.g.* “cause”). Select Total Y to 100% within each X category if you want the bars to represent percentages within each level of the X variable (the one defining the different groups of bars).

Stacked frequencies

The bar charts are made as just described for individual observations, but in the Graph Variables box, set Function to Sum, and Y to the variable containing the frequencies.

ONE-SAMPLE PROCEDURES

t procedures

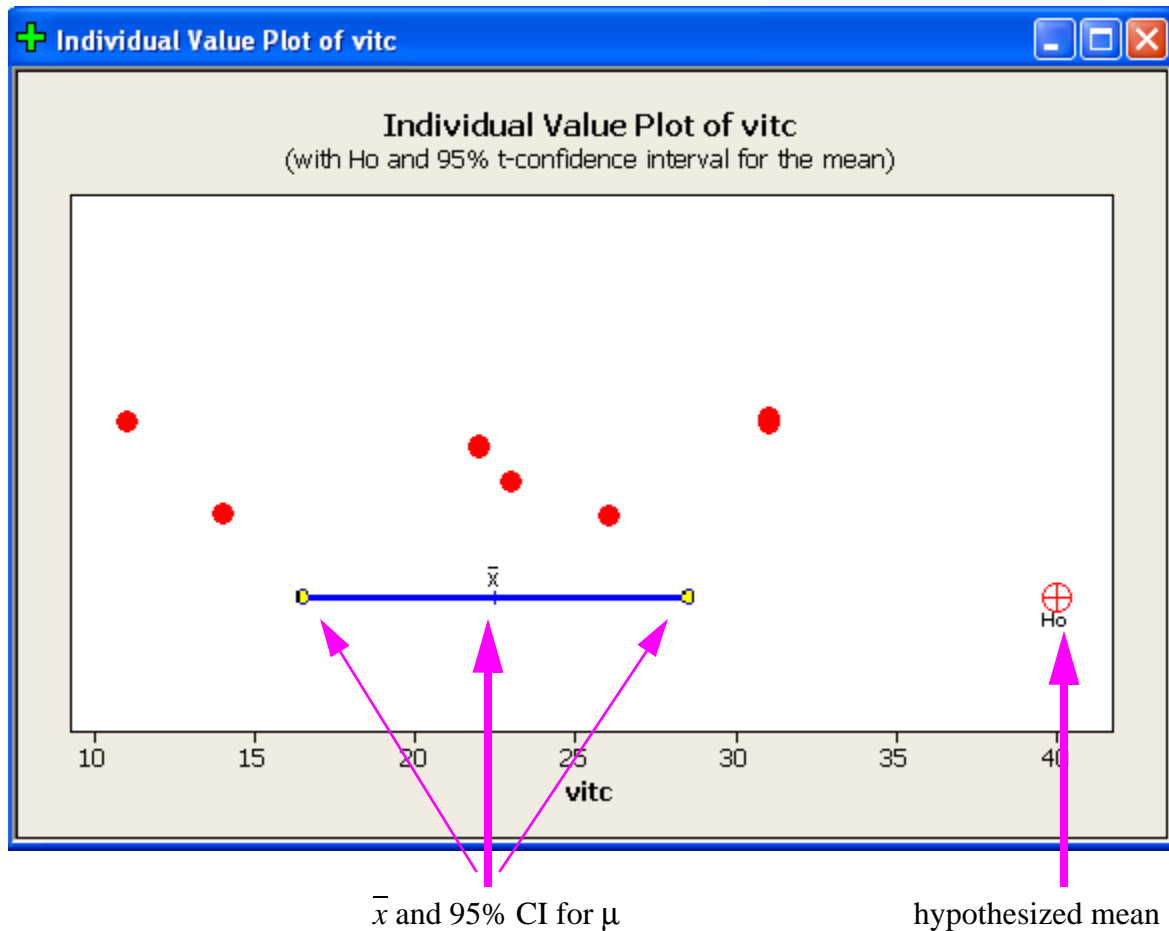
Stat ⇒ **Basic Statistics** ⇒ **1-Sample t...**

To analyze raw data, select **Samples in columns:** and specify the variable(s) to be analyzed. Alternatively, if the sample statistics have already been calculated, select **Summarized data** and enter the **Sample size:**, **sample Mean:**, and **sample Standard deviation:** in the appropriate boxes. If a test is desired, the mean specified by H_0 also must be specified (**Test mean:** box).

The screenshot shows the '1-Sample t (Test and Confidence Interval)' dialog box. On the left, a list of variables is shown: C1 id and C2 vite. The 'Samples in columns:' radio button is selected, and 'vite' is entered in the list box. The 'Summarized data' radio button is unselected. Below it, there are three empty input boxes for 'Sample size:', 'Mean:', and 'Standard deviation:'. The 'Test mean:' box contains the value '40' and is followed by the text '(required for test)'. At the bottom, there are buttons for 'Select', 'Help', 'Graphs...', 'Options...', 'OK', and 'Cancel'.

The **Options...** button opens a dialog in which the confidence level can be changed (from default 95%) and/or a one-sided test can be specified.

The Graphs... button allows histograms, boxplots, and/or individual value plots to be requested. In this, \bar{x} , the CI, and μ_0 are shown along the appropriate axis, as shown in the example below.



Distribution-free procedures

Sign procedures

Stat ⇒ **Nonparametrics** ⇒ **1-Sample Sign...**

Then specify the variable(s) to be analyzed, whether you want a confidence interval or hypothesis test (can't do both at once), and either the confidence level or the value of the median under H_0 and the direction of H_a .

Output

Because the distributions of the test statistics are discrete for most distribution-free procedures, there typically is not a value which would exactly a given confidence level

such as 95%. For all distribution-free CIs except the sign CI Minitab gives the CI for the achievable exact confidence level closest to the level requested.

For sign CIs, however, Minitab provides three CIs, as shown below:

- one for the closest achievable exact confidence level below the level requested (92.97% in this example),
- one for the closest achievable exact confidence level above the level requested (99.22% in this example), and
- one which Minitab interpolates between the previous two, as an approximation for the actual confidence level requested; the latter has NLI in the `Position` column of the output table.

Sign CI: vitc						
Sign confidence interval for median						
	N	Median	Achieved Confidence	Confidence Interval		Position
				Lower	Upper	
vitc	8	22.50	0.9297	14.00	31.00	2
			0.9500	13.81	31.00	NLI
			0.9922	11.00	31.00	1

Signed-rank procedures

Stat ⇒ **Nonparametrics** ⇒ **1-Sample Wilcoxon...**

Then specify everything as described above for the sign procedures.

- Output is the same except no interpolation is performed to get the requested confidence level for the CI; instead the achievable confidence level nearest the requested level is given (here, 94.1%).

Wilcoxon Signed Rank CI: vitc					
	N	Estimated Median	Achieved Confidence	Confidence Interval	
				Lower	Upper
vitc	8	22.5	94.1	16.5	28.5

Resampling procedures

CIs

Confidence intervals can be estimated for the mean, the median, the standard deviation, or any other statistic programmed by the user. Each of these statistics uses a different macro, but the four are very similar, so only the one for the mean will be described in detail here.

Estimation methods

Six kinds of confidence intervals are produced (besting S-Plus for quantity if not quality; the “tilting” method recommended by S-Plus is not available).

- Estimate $\pm 1.96 \cdot \text{boot sd}$

This simple method uses the standard deviation of the bootstrap distribution of the statistic to estimate the standard error of the observed sample statistic, and uses the critical value for a 95% interval assuming a normal (not t) sampling distribution. **Note that this does not change if you request a confidence level other than 95%.**
- Bootstrap- t method

This determines the critical value not from normal or t distributions, but from percentiles of the distribution of t statistics calculated for the bootstrap samples. This is multiplied by the standard error of the statistic based on the observed sample standard deviation. **Note that this is different from the “bootstrap t CI” described in Chapter 14 of the text**, which uses the bootstrap distribution to estimate the standard error and multiplies this by the critical value of the t distribution.
- Efron percentile method

This simply uses the appropriate percentiles of the bootstrap distribution of the statistic as the limits of the CI.
- Hall percentile method

This gives sort of the mirror-image of Efron’s method: the lower limit is as far below the observed value of the statistic as Efron’s upper limit is above it, and vice versa.
- BC percentile method

“**Bias Corrected**”: Efron’s method after correction for possible bias (as estimated from the difference between the observed statistic and the mean of the bootstrap distribution); the correction alters which percentiles of the bootstrap distribution are used.
- BCA percentile method

“**Bias Corrected – Accelerated**”: the BC method with further correction for possible non-constant standard error.

CI for Mean

```
MTB > meanciboot c1
```

The column containing the observations is given in place of **c1**.

Options

```
SUBC> siglev k1
```

confidence level (as %); default is 95

```
SUBC> nboot k1
```

number of bootstrap samples; default is 2000

```
SUBC> means c1
```

column for the bootstrap means

```
SUBC> quantiles c1-c3
```

storage for the ranks (within the bootstrap distribution) of the CI limits; the three columns are for the Efron, BC, and BCa methods.

Output

The output is in three sections. The first gives basic statistics and standard *t* confidence intervals, while the second gives some information about the bootstrapping (e.g. the number of resamples and the mean and standard deviation of the bootstrap means). The third section is the actual confidence intervals:

Estimate +/- 1.96*boot sd	1.667	8.619
Bootstrap-t method	1.010	8.952
Efron percentile method	1.586	8.348
Hall percentile method	1.938	8.700
BC percentile method	1.586	8.367
BCA percentile method	1.538	8.300

These macros generally do not produce any graphical output but by saving the bootstrap distribution (e.g. in **means c1**) you can make any graphs you want, of which the most useful will be a histogram and a NQQ plot. For some reason **stdevciboot** produces such a histogram automatically.

CI for Median

```
MTB > medianciboot c1
```

The column containing the observations is given in place of **c1**. Options are the same as for **meanciboot** except the subcommand for storing the bootstrap distribution of medians is given as **medians c1**.

CI for Standard deviation

```
MTB > stdevciboot c1
```

The column containing the observations is given in place of **c1**. Options are the same as for **meanciboot** except the subcommand for storing the bootstrap distribution of standard deviations is given as **stdevs c1** and the **tvalues** option is not available.

CI for "Any statistic"

The macro file has a few lines which can be modified to change what statistic is bootstrapped. This can be a function or another macro; it must take a single column as input and produce a single value as output. If you want to use this macro, see me for help.

Significance test

There is only one resampling test available for a single sample: a test of whether the mean of the population equals a hypothesized value.

```
MTB > onesampleran c1 k1
```

The column containing the observations is specified in place of **c1** and the population mean specified by the null hypothesis is given in place of **k1**.

Options

```
SUBC> nran k1
```

the number of randomizations; default is 999

```
SUBC> sums c1
```

column to store the sample sums for the randomizations

Method

This test is sort of a hybrid of the signed-rank test and a t test. First the hypothesized mean is subtracted from all observations (as is done with the hypothesized median in a signed-rank test). Then the signs of these adjusted observations are randomly determined, and the (now randomly signed) values are summed. If the null hypothesis is true, the (adjusted) values will be centered close to 0, and so will sum to close to 0, and in addition any given observation has equal chance of being positive or negative.

Output

The output is simply some descriptive statistics and then P -values for both one-sided tests and the two-sided test.

PAIRED-SAMPLE PROCEDURES

Analyzing differences

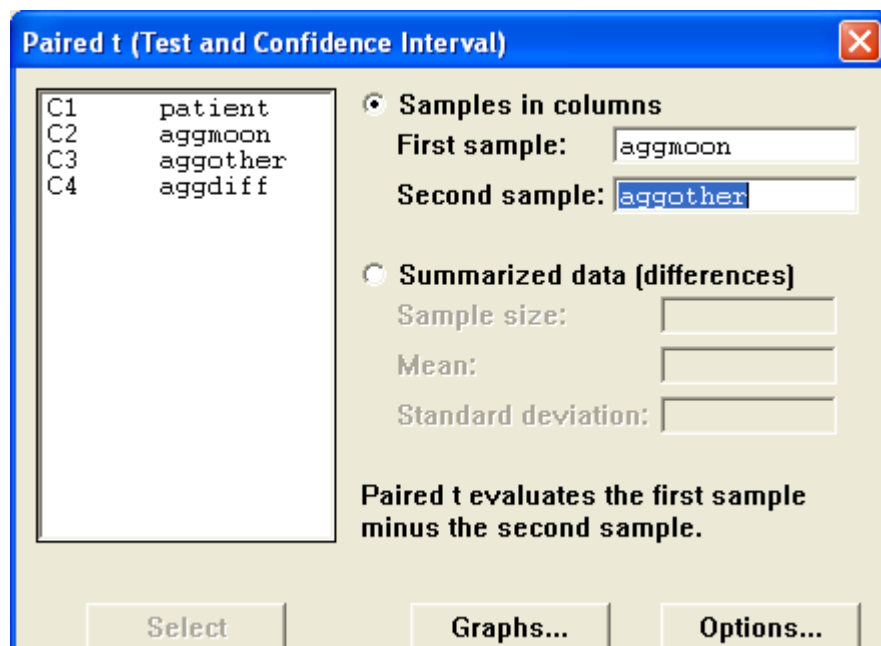
Any of the paired-sample procedures can be implemented by first calculating within-pair differences, using the Transform dialog on the Data menu, and then applying the appropriate one-sample procedure (as described in the previous chapter) to the differences. **The distribution-free tests (sign and signed-rank) and the resampling analyses can only be done this way.**

Note: While these tests usually are applied to within-pair differences, there could be circumstances in which some other within-pair comparison, e.g. a ratio, might be more appropriate. In this case, simply compute the desired within-pair measure and apply the one-sample procedures as usual.

Paired-sample t test

Stat ⇒ **Basic Statistics** ⇒ **Paired t...**

Specify the two columns containing the samples:



Graphs (of differences) and Options are as for a one-sample t test. If summary statistics of the differences have been calculated, the Summarized data option of the Paired t can be used.

TWO-SAMPLE PROCEDURES

t procedures

Stat ⇒ **Basic Statistics** ⇒ **2-Sample t...**

Unstacked data

If the data are unstacked (each sample in a separate column, as in the example data set to the right), select **Samples in different columns**, and specify the two columns of data in the **First:** and **Second:** boxes; which is which matters only for interpreting the direction of the difference in means.

C6	C7
adults	juveniles
0.00000	0.333333
	* 0.500000
0.00000	0.500000
0.23077	0.500000
0.00000	0.105263
0.00000	0.538462
0.00000	0.666667

2-Sample t (Test and Confidence Interval)

C1 sexage C2 totfood C3 algdet C4 propinvert C6 adults C7 juveniles C8 C9 freq	<input type="radio"/> Samples in one column Samples: <input type="text"/> Subscripts: <input type="text"/>
	<input checked="" type="radio"/> Samples in different columns First: <input type="text" value="adults"/> Second: <input type="text" value="juveniles"/>

Stacked data

If the data are stacked (the analysis variable in one column and a grouping variable in another column, as in the example to the right), select **Samples in one column**, enter the quantitative (response) variable in the **Samples:** box, and the categorical variable defining the two samples in the **Subscripts:** box.

C4	C5-T
propinvert	age1
0.00000	adult
	* adult
0.00000	adult
0.23077	adult
0.00000	adult
0.00000	adult
0.33333	juvenile
0.00000	adult
0.03704	adult
0.00000	adult
0.50000	juvenile

2-Sample t (Test and Confidence Interval)

C1 sexage C2 totfood C3 algdet C4 propinvert C5 age1 C6 adults C7 juveniles C8 C9 freq C10 age C11 side	<input checked="" type="radio"/> Samples in one column Samples: <input type="text" value="propinvert"/> Subscripts: <input type="text" value="age1"/>
	<input type="radio"/> Samples in different columns First: <input type="text"/> Second: <input type="text"/>

If you already have the sample sizes, sample means, and sample standard deviations, select Summarized data and enter the values of these statistics for the two samples.

If desired, check the checkbox to Assume equal variances. The confidence level, the difference in means specified by H_0 , and a one-sided test can be changed from the defaults (95%, 0 difference, two-sided) from the Options... button.

Available graphs (from the Graphs... button) are side-by-side boxplots and/or dot-plots, with the sample means indicated.

Resampling procedures

The macros

Four resampling macros are applicable for comparing two independent samples:

- `twosampleran`
performs a randomization test of the null hypothesis that the two population distributions from which the samples were obtained are identical; requires the samples to be in separate columns (unstacked).
- `twotran`
performs the same test as `twosampleran` but requires the data to be in stacked layout (one column with the quantitative variable to be analyzed, and another column — which must be numeric — identifying the two samples).
- `twotunpoolboot`
uses bootstrapping to perform a test of the hypothesis of equal population means, without assuming equal variances (i.e. not using a pooled estimate of the variance).
- `twotpoolboot`
also uses bootstrapping to test the hypothesis of equal population means, but pools the samples to estimate the variance and thus assumes equal variances.

The four two-sample macros all test the hypothesis that two population means are equal. Two use a randomization test and the other two use a bootstrap test. All four are very similar in how they are specified and in their output, so only the first will be described in detail.

Randomization tests

The randomization in the first two tests consists of randomly re-allocating observations to the two samples (while maintaining the observed sample sizes). **Note that this implies the null hypothesis that the population distributions are identical, including having the same spread.**

Unstacked data

```
MTB > twosampleran c1 c2
```

The two samples are in the columns specified in the command (in place of `c1` and `c2` above).

Options

```
SUBC> nran k1
```

the number of randomizations; default is 999.

```
SUBC> differences c1
```

column to store the between-sample differences for the randomizations

```
SUBC> tstatistics c1
```

column to store the t statistics for the randomizations

Output

The output is some descriptive statistics and then P -values for both one-sided tests and the two-sided test.

Stacked data

```
MTB > twotran c1 c2
```

In this case the column containing the observations is specified first (in place of `c1`) and the column identifying which sample an observation is from is specified second (in place of `c2`). **This group variable must be numeric.**

The options and output are identical to those for `twosampleran` above.

Bootstrap tests

The other two two-sample tests take resamples (i.e. with replacement), rather than shuffling the observations among the two samples. Before the resampling, the respective sample means are subtracted from the observations. How the resampling is done depends on whether the variances of the populations are assumed to be equal, as described below.

Data for either of these bootstrap test macros must be stacked.

Unpooled

```
MTB > twotunpoolboot c1 c2
```

The column containing the observations is specified first (in place of `c1`) and the column identifying which sample an observation is from is specified second (in place of `c2`). **The group variable must be numeric.**

The options and output are the same as for the two randomization tests above, except that the subcommand setting the number of resamples is given as **nboot k1** rather than **nran k1**.

Method

This procedure does not assume equal variances and therefore does not pool the samples in any way. First each sample is centered by subtracting off its mean; this results in two samples both with mean 0 but possibly different spread (and shape, for that matter). Then in each iteration of the bootstrapping, each sample is resampled (i.e. the same number of observations is randomly sampled, with replacement), and the unpooled-*t* statistic for the two samples is computed.

Pooled

```
MTB > twotpoolboot c1 c2
```

Specification, options, and output are the same as for **twotunpoolboot**.

Method

This procedure does assume equal variances and does pool the samples in the resampling. First each sample is centered by subtracting off its mean; this results in two samples both with mean 0. These then are pooled. In each iteration, n_1 observations are resampled (with replacement) from the pooled set of (adjusted) observations and assigned to sample 1, n_2 are resampled and assigned to sample 2, and the pooled *t* statistic then is computed for these two reconstituted samples. (Here n_1 and n_2 are the sizes of the observed samples.)

Distribution-free procedures

Rank-sum procedures

Unstacked data

If the data are unstacked (each sample in a separate column, as in the upper example on the first page of this chapter), use

Stat ⇒ **Nonparametrics** ⇒ **Mann-Whitney...**

Then specify the columns containing the data (see first dialog window on next page). Change the confidence level and/or choose a one-sided test if desired.

Stacked data

If the data are stacked, use

Stat ⇒ **Nonparametrics** ⇒ **Kruskal-Wallis...**

Then specify the quantitative variable as the **Response:** and the categorical variable defining the two samples as the **Factor:** (see second dialog window below).

Mann-Whitney

C1	sexage
C2	totfood
C3	algdet
C4	propinvert
C6	adults
C7	juveniles
C8	
C9	freq

First Sample: adults

Second Sample: juveniles

Confidence level: 95.0

Alternative: not equal

Kruskal-Wallis

C1	sexage
C2	totfood
C3	algdet
C4	propinvert
C5	age1
C6	adults
C7	juveniles
C8	

Response: propinvert

Factor: age1

Median test

Data must be in stacked form: observations in one column, subscripts in another.

Stat ⇒ **Nonparametrics** ⇒ **Mood's Median Test...**

Then specify the quantitative variable as the Response: and the categorical variable defining the samples as the Factor:. If desired, check the boxes to store residuals and fits.

Mood's Median Test

C1	sexage
C2	totfood
C3	algdet
C4	propinvert
C5	age1
C6	adults
C7	juveniles
C8	
C9	freq
C10	age
C11	side

Response: propinvert

Factor: age1

Store residuals

Store fits

SEVERAL-SAMPLE PROCEDURES

ANOVA

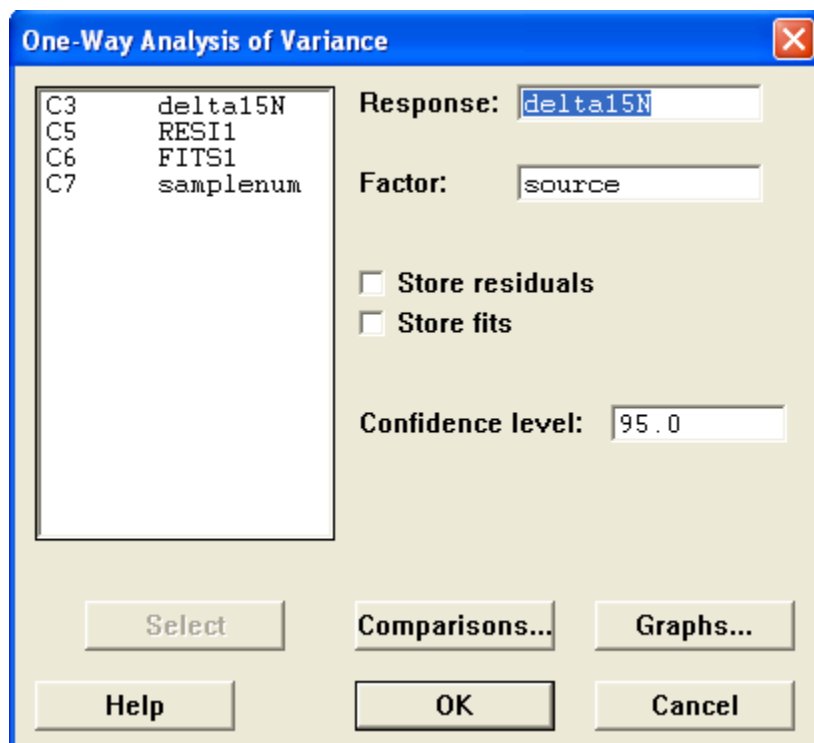
Dialog

Stacked data

If the data are stacked, use

Stat ⇒ **ANOVA** ⇒ **Oneway...**

When specifying the data, the “Response:” column is the one with the observations (the quantitative variable), and the “Factor:” column is the one with the categorical variable identifying the samples.

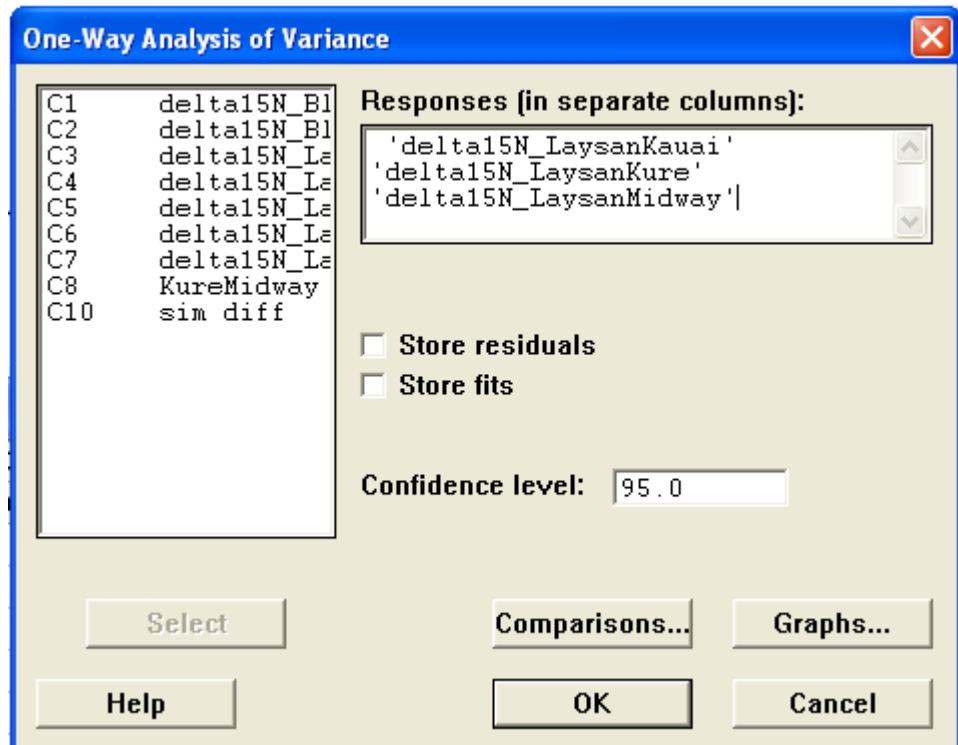


Unstacked data

If the data are unstacked, use

Stat ⇒ **ANOVA** ⇒ **Oneway (Unstacked)...**

In this case, the columns containing the variables for the different samples are specified in the Responses (in separate columns): box; there is no “factor” variable in this layout.



Options

In either version of the procedure, check boxes allow you to Store residuals and/or Store fits.

The output includes CIs for the means of the groups, based on the pooled standard deviation and with no adjustment for multiple inferences; the Confidence level: that can be changed in the main ANOVA window is for these CIs.

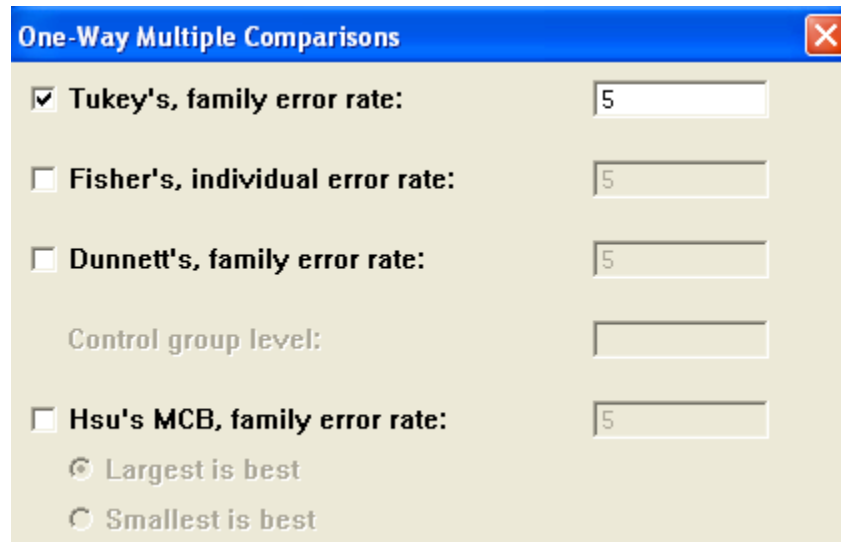
Multiple comparisons

Clicking the Comparison... button opens a window (see below) in which one or more methods of multiple comparisons can be selected by checking the appropriate boxes. The most useful of these methods are:

- Tukey’s (the best of these for comparing every group to every other group)
- Dunnett’s (for comparing one “reference” or “control” group to each of the other groups)

(“Hsu’s MCB” is for determining which groups are “better” than which others, where “better” you define either as “larger than” or as “smaller than”. Fisher’s compares every group to every other group, with no adjustment for making multiple comparisons.)

The default α for these comparisons is 0.05; if a different levels is desired it can be entered in the appropriate box. Any error rate greater than 1 will be interpreted as a



percent, while any value less than 1 will be interpreted as a proportion; the error rate must be between 0.5 (50%) and 0.001. (Note that for Fisher's comparisons this α is the error rate for each individual comparison, while for the other methods it is the “family” error rate.)

Residual and other plots

The Graphs... button opens a window in which several kinds of plots can be selected. Individual value plot (a dot plot) and Boxplots of data are plots of the actual samples.

Several standard Residual Plots can be requested, as Individual plots or as Four in one (stacked data) or Three in one (unstacked data) plots which put all the available residual plots in one graph window:

- Histogram of residuals
- Normal plot of residuals
- Residuals versus fits, a plot of the residuals against the sample means (the “fits”)
- Residuals versus order [available only in the stacked form], a plot of the residuals against their order in the data set.

In the stacked layout, the residuals can be plotted against any other quantitative variable(s) in the data set, by specifying them in the Residuals versus the variables: box.

Output

Standard output

The default output includes:

- a standard ANOVA table
- a line giving the square root of MSR (“S”), R^2 (“R-Sq”), and the “adjusted R^2 ” (“R-Sq(adj)”; this measure is not useful for one-way ANOVA, so ignore it)

One-way ANOVA: delta15N versus source

Source	DF	SS	MS	F	P
source	2	5.463	2.731	5.33	0.009
Error	39	19.979	0.512		
Total	41	25.441			

S = 0.7157 R-Sq = 21.47% R-Sq(adj) = 17.44%

Individual 95% CIs For Mean Based on
Pooled StDev

Level	N	Mean	StDev	CI Lower	CI Upper
Kauai	26	13.032	0.691	12.40	13.20
kure	12	12.245	0.665	12.00	12.80
midway	4	12.463	1.029	11.50	13.50

Pooled StDev = 0.716

- a table giving descriptive statistics for each group and a crude chart of CIs for the group means (as noted above, these are without any adjustment for multiple comparisons), and
- the pooled estimate of the standard deviation of the “error” term (“Pooled StDev”, which is exactly the same, apart from rounding, as the “S” given earlier in the output).

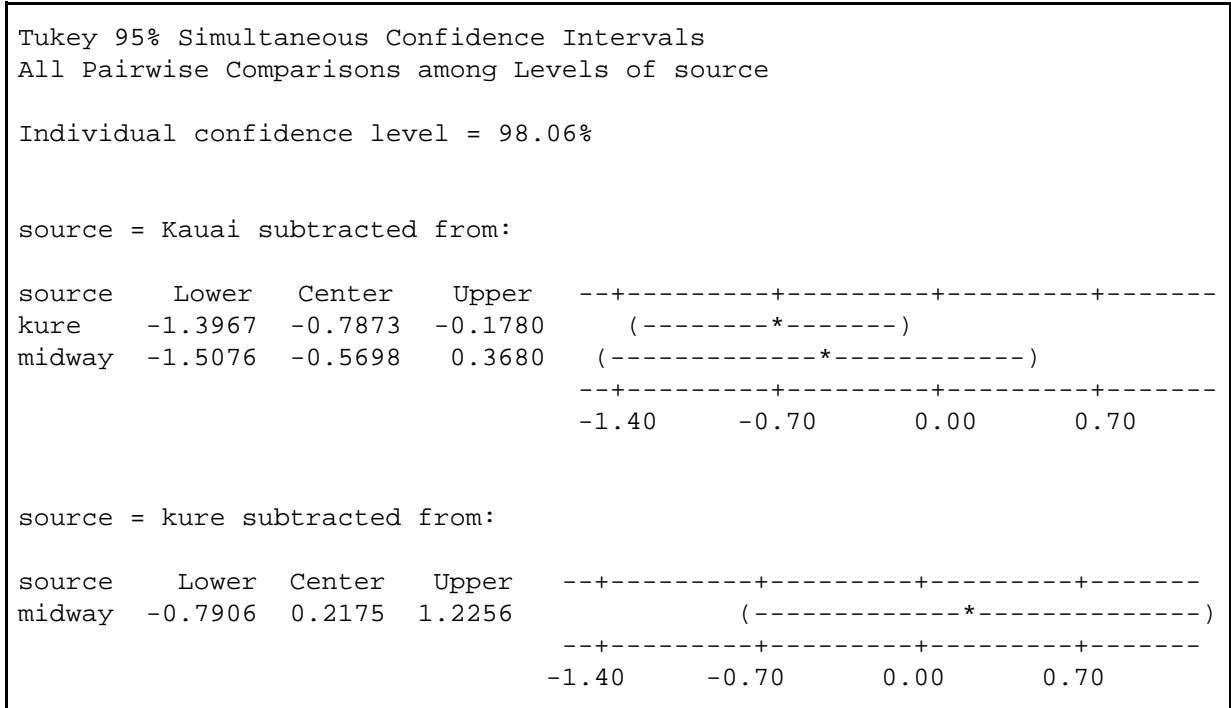
Multiple comparisons

If multiple comparisons are requested, their output is given after the default output described above (see example below). After a heading saying what kind of comparisons they are and at that confidence level, there is a line stating the “test-wise” confidence level corresponding to the chosen “family-wise” confidence level.

Then come the actual results, in the form of a series of tables, each comparing one group to each of the subsequent groups. The first table compares the first group to each of the others, the second table compares the second group to all remaining groups (i.e. all but the first), etc. (Group ordering is alphabetical by the group labels.) Each table is labelled “*factor* = *value* subtracted from:” where the name of the grouping variable (the ANOVA factor) takes the place of *factor* and the label of the reference group for that table takes the place of *value*. The rows in the table then are labelled (in the column headed “*factor*”) by the label of the group being compared to that table’s reference group. The pairwise comparisons in each row are given as

- the estimated difference in population means (estimated simply by the difference in sample means), listed in the column headed “Center”, surrounded by

- the CI for the difference in population means, listed in the columns headed “Lower” and “Upper”, followed by
- a crude chart of these estimates.



In the example here, the first table of comparisons contains the two comparisons, Kauai vs. kure and Kauai vs. midway. The first row this first table indicates that the mean of the kure sample was 0.7873 units smaller than that of the Kauai sample, with a CI of -1.3967 to -0.1780 for the difference between the means of the populations (both the difference and the CI being expressed in terms of the (kure - Kauai) difference). The second row of the first table then gives the midway - Kauai comparison. The second (and last) table then gives the one remaining comparison, midway vs. kure. In this example, the Kauai and kure samples were significantly different at the (family-wise) 95% level, since the CI did not contain 0, but neither of these was significantly different from the midway sample.

Test for equal variances

Data must be stacked.

Stat ⇒ **ANOVA** ⇒ **Test for Equal Variances...**

Specify the “Response:” and “Factor:” as for an ANOVA.

Output

This test produces both text (tabular) and graphical output. The former includes estimates and confidence intervals for the population standard deviations, and results of

two tests of the null hypothesis that all population variances are equal. Of these, Levene's test is more reliable. A large P -value, as in this example, is support for the ANOVA assumption of equal variances.

```

Test for Equal Variances: delta15N versus source

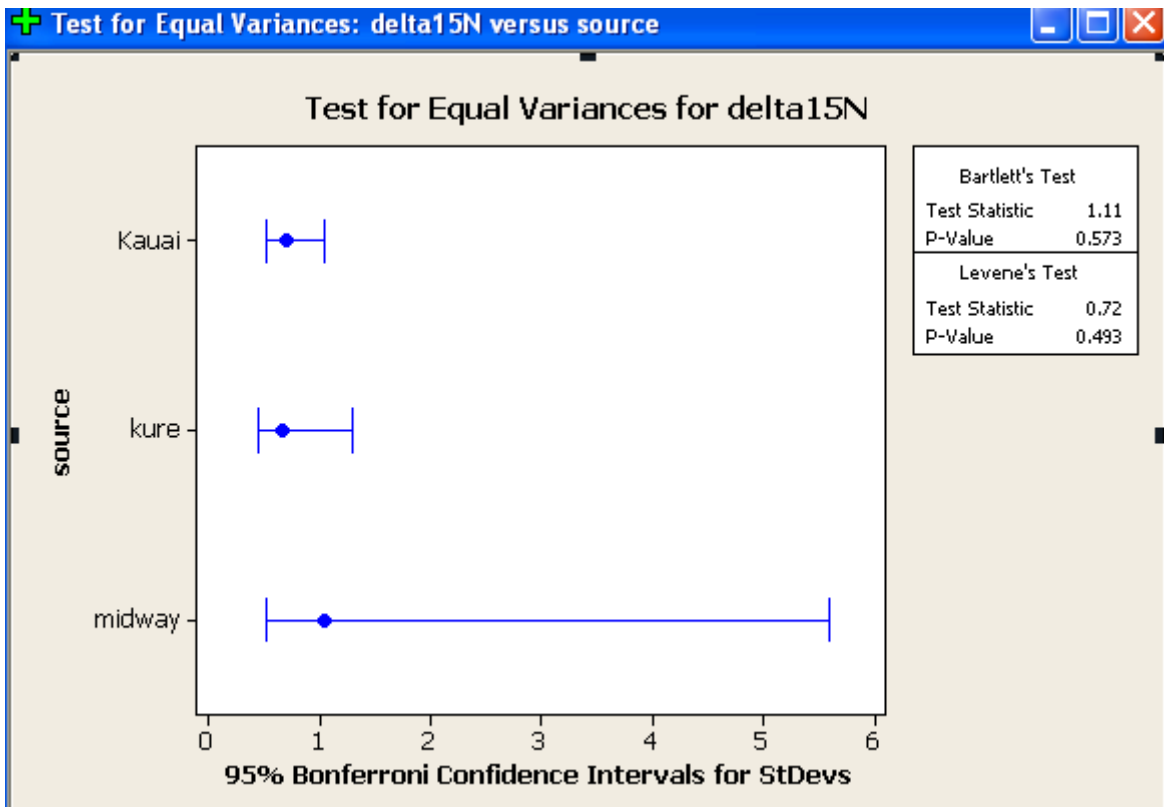
95% Bonferroni confidence intervals for standard deviations

source    N      Lower    StDev    Upper
Kauai    26    0.514836  0.69083  1.03004
  kure    12    0.439100  0.66548  1.29028
midway    4     0.520154  1.02893  5.59612

Bartlett's Test (normal distribution)
Test statistic = 1.11, p-value = 0.573

Levene's Test (any continuous distribution)
Test statistic = 0.72, p-value = 0.493
    
```

The graphical output simply graphs the estimates and CIs of the population standard deviations, and gives the test results in boxes to the side.



Resampling procedures

The macros

Two resampling macros are applicable to comparing two or more samples:

- **onewayran**
performs a randomization test of the null hypothesis that all population means are equal.
- **leveneran**
performs a bootstrap version of Levene's test for equal variances.

ANOVA

```
MTB > onewayran c1, c2
```

The data must be in stacked layout, with the analysis variable listed first (i.e. in place of *c1*) and then the variable defining the groups (in place of *c2*). The group variable must be numeric.

Options

```
SUBC> nran k1
```

the number of randomizations; default is 999.

```
SUBC> fvalues c1
```

a column for storage of the *F* ratios from the randomizations

No preplanned contrasts or unplanned multiple comparisons are available.

Method

This is an extension of the randomization test for two samples: the observations are randomly shuffled among the groups (maintaining the observed sample sizes). The test statistic, calculated for each randomization, is the *F* statistic (MS Groups / MS Error). The observed *F* statistic is compared to the distribution of *F* statistics from the randomizations to determine the *P*-value.

Output

This macro first produces the output from the standard one-way ANOVA (ANOVA table and descriptive statistics for the groups). The only new output is simply the randomization *P*-value.

Test for constant variance

```
MTB > leveneran c1, c2
```

The data must be in stacked layout, with the analysis variable listed first (i.e. in place of `c1`) and then the variable defining the groups (in place of `c2`). The group variable must be numeric.

Options

SUBC> nran <i>k1</i>	the number of randomizations; default is 999.
SUBC> fvalues <i>c1</i>	a column for storage of the <i>F</i> ratios from the randomizations
SUBC> usemean <i>k1</i>	a flag to have the test use deviations from group <u>medians</u> (<i>k1</i> = 0, the default) or group <u>means</u> (<i>k1</i> = 1).

Method

This is a version of Levene's test with the *P*-value determined by randomization. The test, for each randomization, is a one-way ANOVA on the absolute values of deviations from the group medians (or from the group means, if the subcommand `usemean 1` is given). The randomization is as in `onewayran` above: observations are shuffled among the groups. The test statistic for each randomization is the *F* statistic from the Levene's test.

Output

The output includes the Levene's test for the observed data, i.e. the ANOVA on the absolute deviations from the group medians. This is followed by the *P*-value determined from the randomizations.

Distribution-free procedures

Kruskal–Wallis test

The data must be in the stacked layout.

Stat ⇒ **Nonparametrics** ⇒ **Kruskal-Wallis...**

Specify the columns containing the Response: (the quantitative variable) and the Factor: (the categorical variable defining the groups). No options or graphs are available.

Output

The output from the Kruskal-Wallis test is minimal:

- a table giving the sample size, sample median, average of ranks, and Z score (standardized average of ranks) for each sample, and

- lines giving the test statistic, its degrees of freedom, and the P -value, first without and then with an adjustment for ties.

```

Kruskal-Wallis Test: delta15N versus source

```

The P -value for this test is gotten by a normal approximation which assumes large sample sizes. The output therefore also sometimes, as in this example, includes a warning about one or more samples being small; the accuracy of the P -value is questionable in these situations.

Median test

The data must be in the stacked layout.

Stat ⇒ **Nonparametrics** ⇒ **Mgod's Median Test ...**

Specify the columns containing the Response: (the quantitative variable) and the Factor: (the categorical variable defining the groups). No graphs are available; the only options are to Store residuals and (if residuals are being stored) to Store fits.

Output

The output from the median test also is minimal:

- the test statistic, degrees of freedom, and P -value
- a table giving, for each group, the number of observations less than the overall median, the number greater than the overall median, the sample median and IQR, and a crude chart of CIs for the population medians,
- the overall median.

Because the CIs for the medians of the different populations are based on sign tests, they cannot attain 95% confidence for groups with few observations; when this occurs the output includes a note to this effect, as in this example here.

Mood Median Test: delta15N versus source

Mood median test for delta15N

Chi-Square = 7.79 DF = 2 **P = 0.020**

source	N<=	N>	Median	Q3-Q1	Individual 95.0% CIs
Kauai	9	17	12.95	1.08	(-----*-----)
kure	10	2	12.01	0.93	(---*-----)
midway	2	2	12.23	1.91	(-----*-----)

-----+-----+-----+-----
 12.00 12.60 13.20

Overall median = 12.70

* NOTE * Levels with < 6 observations have confidence < 95.0%

REGRESSION AND CORRELATION

Linear least-squares regression

Command

Stat ⇒ **Regression** ⇒ **Regression...**

Then specify the columns containing the data:

- the “Response:” column is the one with the response variable,
- the “Predictors:” column is the one (or more) with the explanatory variable(s)

Options

The only option relevant for this course is “Prediction intervals for new observations:” (which includes confidence intervals for the mean response). For this option, click the Option button on the regression dialog window and then

- specify value(s) of the explanatory variable for which this estimation is desired; this can be done by specifying a column containing the set of values,
- change the confidence level if desired, and
- choose which if any of these quantities should be stored.

Graphs

There are various residual plots available. These can use one of several types of residuals:

- “Regular” residuals:
simply the difference between the observed and “fitted” values
- “Standardized” residuals
regular residuals divided by their standard deviations.
- “Deleted” residuals
for more advanced purposes, discussed in Advanced Biometry.

These residuals can be plotted either as Individual plots, or as Four in one plots which put all the available residual plots in one graph window. As in ANOVA, the plots are:

- Histogram of residuals
- Normal plot of residuals
- Residuals versus fits, a plot of the residuals against the sample means (the “fits”)
- Residuals versus order, a plot of the residuals against their order in the data set.

The residuals can be plotted against any other quantitative variable(s) in the data set, by specifying them in the Residuals versus the variables: box.

Storage

If desired, various quantities resulting from the analyses can be stored; of these, the ones relevant this semester are:

- coefficients (intercept and slope),
- Fits (estimated values)
- Residuals, and/or
- Standardized residuals.

Correlation (Pearson's)

Stat ⇒ **Basic Statistics** ⇒ **Correlation**

then specify the two (or more) variables for which you want correlations; you can have the correlation matrix saved rather than printed if you wish.

The P -value is for the test of $H_0: \rho = 0$ vs. $H_a: \rho \neq 0$; any other inference about ρ must be done by hand.

Resampling procedures

The macros

The applicable resampling macros are:

- `regresssimran`
[note the 3 successive s'es] performs a randomization test of the null hypothesis that the regression slope is 0.
- `correlationran`
performs a randomization test of the null hypothesis of 0 correlation. (This test is identical to that performed by `regresssim` but the output is in terms of the correlation coefficient rather than a regression equation.)

Internally these two macros are the same, but the printed output and storage options differ according to their different purposes. Both test the null hypothesis that the two variables are unrelated, expressing this either as $H_0: \beta_1 = 0$ or $H_0: \rho = 0$.

Simple linear regression

```
MTB > regresssimran c1 c2
```

The response variable is listed first (in place of **c1**), followed by the explanatory variable (in place of **c2**). Note that there are 3 consecutive s'es in the command.

Options

SUBC> nran k1	the number of randomizations; default is 999.
SUBC> fits c1	a column for storage of the fitted values <u>from the original regression on the observed data</u> .
SUBC> residuals c1	a column for storage of the residuals <u>from the original regression on the observed data</u> .
SUBC> correlations c1	a column for storage of the correlation coefficients from the randomizations.
SUBC> coefficients c1 c2	two columns, for storage of intercept estimates (in c1) and slope estimates (in c2) from the randomizations.
SUBC> tstatistics c1	a column for storage of the <i>t</i> statistics from the randomizations.

Method

The values of the response variable are randomly shuffled among the observations, with the values of the explanatory variable kept constant.

Output

The output is quite minimal: after a few descriptive statistics, including the estimated regression coefficients, randomization *P*-values are given for the two-sided and both one-sided tests.

Correlation

```
MTB > correlationran c1 c2
```

In this case the order in which the variables are listed obviously doesn't matter.

Options

SUBC> nran k1	the number of randomizations; default is 999.
SUBC> corrs c1	a column for storage of the correlation coefficients from the randomizations.

Method

Analogous to the method of `regresssimran` above: the values of the second variable are randomly shuffled among the observations, with the values of the first variable kept constant.

Output

As with `regresssimran`, the output includes a few descriptive statistics, including the estimated correlation coefficient, and randomization P -values for the two-sided and both one-sided tests.

Distribution-free procedures

Spearman's rank correlation

First transform the data to ranks, then calculate the correlation coefficient as above (or run a regression). To rank the data:

Data \Rightarrow **Rank** ...

then specify the column containing one of the variables (Rank data in:) and a column to Store ranks in; note that only one variable can be done at a time.

Then obtain the correlation coefficient as above for Pearson's correlation, but using the columns containing the ranks. The P -value can be considered approximately correct for the Spearman's correlation.

Kendall's τ (tau)

There is no direct procedure for Kendall's procedure, but there is a built-in command which facilitates hand-calculating it by calculating the slopes between all pairs of observations. The number of these which are positive, and the number which are negative, can then be counted by using true-false comparisons and column summation in the calculator; these counts then can be used to calculate the estimate of τ . Inference must be done by hand.

1. **Stat** \Rightarrow **Nonparametrics** \Rightarrow **Pairwise Slopes**
and specify Y and X variables, and a column to store the slopes in.
2. **Calc** \Rightarrow **Calculator**
and specify a column to store the results (`true/false` (1/0) values) in, and specify the expression `Cs > 1` (C_s is the column of pairwise slopes created above).
3. **Calc** \Rightarrow **Column Statistics** ...
and specify the column created in the previous step as the "Input variable:". This result will be the count of positive slopes (K'). To get the count of negative slopes (K'') repeat the two **Calc** steps but give the expression in the first step as `Cs < 1`.
4. These counts (K' and K'') can then be used to estimate τ and/or get a P -value by hand.

CATEGORICAL VARIABLES: PROPORTIONS

This section deals with relationships between two categorical variables, that is, comparison of two or more proportions or conditional distributions, as represented in a two-way contingency table. Inference about a single proportion, whether hypothesis testing or interval estimation, and whether using the normal approximation or exact binomial probabilities, must largely be done by hand, though the probability functions described in the next section can help with this.

Data layouts

As described in the section **Describing Relationships**, data for a two-way contingency table can appear in the worksheet in three different layouts. What can be done with the data, as well as how it is done, depends on the layout.

- **Individual observations**
Every independent observation can be in a separate row, with two (or more) columns containing the categorical variables by which each observation is cross-classified.
- **Stacked frequencies**
Each combination of levels of the categorical variables can be in a separate row, with columns containing the categorical variables defining these combinations, and an additional column containing the frequencies of the category combinations.
- **Unstacked frequencies**
The row-and-column layout of a contingency table can be directly represented in the rows and columns of the worksheet, with separate columns containing different levels of one of the categorical variables and separate rows representing different levels of the other variable:

Inference (chi-square test)

Individual observations

Stat ⇒ **Tables** ⇒ **Cross Tabulation...**

then specify:

- the classification variables (note that there can be more than two of these)
- what statistics you want displayed: counts; row, column, and/or total percents; expected counts; standardized residuals (“ X^2 contributions”).

The “Options” button opens up a window allowing other details of what statistics are printed to be altered.

The “Summaries” button opens up a window allowing you to request various descriptive statistics for quantitative variables; these will be calculated and reported for each combination of values of the classification variables.

Stacked frequencies

As for individual observations above, but check the “**Frequencies are in...**” box and specify the column containing the frequencies.

Unstacked frequencies

Stat ⇒ **Tables** ⇒ **Chi-Square Test...**

and specify the columns containing the table of frequencies.

POWER ANALYSIS AND MISCELLANEOUS PROBABILITY PROCEDURES

Power/sample size analysis

Commands

- Stat ⇒ Power and Sample Size ⇒ 1-Sample t...**
- Stat ⇒ Power and Sample Size ⇒ 2-Sample t...**
- Stat ⇒ Power and Sample Size ⇒ 1 Proportion...**
- Stat ⇒ Power and Sample Size ⇒ 2 Proportions...**
- Stat ⇒ Power and Sample Size ⇒ One-way ANOVA...**

All these require you to specify two out of three pieces of information:

- (1) the sample size(s);
- (2) the magnitude(s) of effect; and
- (3) the power value(s).

For planning a study ordinarily you would specify effect size and desired power, to find the required sample size. Specifying sample size and effect size lets you determine the power, perhaps of an already completed study. Similarly, specifying sample size and desired power lets you determine how small an effect could have been “detected.”

More than one value can be specified for any of these quantities, allowing you to easily explore a range of possibilities.

Sample sizes

For two- or multi-sample tests, it is assumed that all samples are of the same size. (Thus if you specify more than one sample size these are treated as different “what if” situations.)

Effect sizes

The hypothesized “effect” is defined differently for the various tests.

- 1-Sample t:
The difference between μ_0 and μ_a .
- 2-Sample t:
The difference between μ_1 and μ_2

- 1 Proportion:
You specify the Hypothesized p (i.e. P_0) and Alternative values of p .
- 2 Proportions:
You specify the two population proportions (with possibly a range of values for one of them).
- One-Way ANOVA:
The difference between the largest and smallest population means. You also have to specify how many levels there are (i.e. how many samples).

Power values

These are entered as probabilities.

Variability

For the t tests and ANOVA you must specify the (expected) standard deviation. It is assumed to be the same for all populations. No measure of variability is needed for the tests of proportions since the variability is determined by the values of p .

Options

For all the analyses, you can change the significance level (i.e. α); its default value is 0.05.

For all the analyses, you also can request that the results be stored. For the one- and two-sample tests you can specify one-sided tests.

Random numbers

Randomizing

Minitab's random-number generator can be used to select a random sample or to randomize treatment over units. The most convenient way to do this is to create a column containing labels of all the units in the population to be sampled or all the experimental units. Then use

Calc \Rightarrow **Random Data** \Rightarrow **Sample From Columns...**

Name the column containing the list of units in the from column(s): box, and another column to store the samples in.

For a random sample, the sample size goes in the Sample _____ rows box. To randomize treatments in an experiment, I find it easiest to set this number equal to the number of units. The result is simply all the units in random order. The first n rows get treatment 1, the next n get treatment 2, etc.

(There is also a check box for sampling with replacement, but this is not appropriate for selecting samples.)

Simulations

The random-number generator also can be used to simulate samples from a wide variety of distributions.

Calc ⇒ **Random Data** ⇒ [*name of distribution*]...

Specify how many observations (rows) to generate, where to put them, and the parameters of the distribution.

Probabilities

Calc ⇒ **Probability Distributions** ⇒ [*name of distribution*]...

You specify the parameters of the distribution, and either specify a column containing the value(s) of interest in the Input column: box or directly enter the values in the Input constant: box. You can optionally specify where to store the results.

This procedure, for any of the many distributions, has radio buttons to choose among:

- **Probability density**
For discrete variables (e.g. binomial distribution) this is the probability of the given value x , as in Table C in the text. For continuous variables (e.g. normal distribution) it is the height of the density curve at the given value; this is not likely to be useful.
- **Cumulative probability**
This is the probability of all values of X less than or equal to the given value x , i.e. the area under the density curve to the left of the value, as in Table A in the text. For discrete values it does include the probability of the exact value. This is the default selection.
- **Inverse cumulative probability**
This is a “critical value,” i.e. the value x which has cumulative probability equal to the value you specify. This is as in Tables D, E, and F in the text, except that it is left-sided (i.e. equal to 1 minus the values in the text tables).

APPENDIX: RESAMPLING MACROS

Macros in Minitab

Macros

Macros in Minitab are text files containing Minitab commands, in a particular format. They can be created by users; if you want to do that, see the Minitab help or other documentation. Some macros are included with Minitab in its standard installation and while primarily used by some of the procedures invoked by the menu interface, presumably they could also be invoked directly from the command line.

Location

How a macro is invoked depends on where the file is located. They can be anywhere in your computer's directory, but by far the most convenient location is `c:\Program Files\MINITAB 14\Macros`. Macros in this folder can be invoked (as described below) while running data from any folder.

Running a macro

Minitab macros are invoked from the command line in the session window. The command line therefore must be enabled. To do this, click in the Session window (to make it the active window), then use the Editor menu and click on Enable Commands.

Once commands are enabled, invoke a macro by typing a `%` sign followed (with no space) by the macro name. For example:

```
MTB > %twosampleran c1 c2
```

If the macro file has the standard extension of `.mac` you do not need to include the extension; if the extension is something else, include it. If the macro file is in the active Windows directory — the one from which you most recently opened a Minitab project or worksheet — or in the Macros folder as given in **Location** above, you do not need to specify the path. If the macro file is anywhere else, include the full path in the macro name.

Required input

All the macros require that you specify the column(s) containing the data, as part of the command invoking the macro. In the example above, `c1` and `c2` are the two columns containing the two samples. **You would of course substitute the proper column names or numbers if the data are not actually in the first two columns.** The *Overview.doc* and *Refcard.doc* references indicate the required input by listing it directly after the macro name.

Options

All the macros accept subcommands to alter options. An option common to them all controls the number of bootstrap samples or randomizations. Most macros allow the test statistic to be stored. Some macros have other specific options.

To enter an option, you **must** end the first macro command line with a semicolon. Minitab then will return not with the usual MTB> prompt but instead with SUBC> . You enter options one at a time (in any order), ending all but the last with semicolons; the last subcommand ends with a period. For example:

```
MTB > %twosampleran c1 c2 ;  
SUBC> nran k1;  
SUBC> differences c1.
```

The first subcommand sets the number of randomization to the value of **k1** (which would be entered as an actual number, e.g. **9999**). The second subcommand names a column to store the resample differences in; this would of course be a different column than the ones containing the observations.

The Resampling Macros

Source

Download the macros as a compressed zip file from

<http://www.ceh.ac.uk/products/software/minitab/download.asp>

Within the file that is downloaded are separate zip files for each of the macros. These can be unzipped individually as needed.

Installation

Unzip the files and put them in an appropriate location, either the

c:\Program Files\MINITAB 14\Macros

folder or a folder where you keep your Minitab data.

Documentation

The downloaded file containing the macros also contains two MS Word documents:

- *Overview.doc* is the complete manual.
- *Refcard.doc* summarizes the commands and subcommands for the macros.

Each macro's folder also contains the manual section for that macro and an example data set.