Genstat®

Introduction

www.vsni.com

VSNi

Introduction to Genstat® *for Windows*™ (21st Edition)

by David Baird, Darren Murray, Roger Payne & Duncan Soutar.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Published by: VSN International, 2 Amberside, Wood Lane,
Hemel Hempstead, Hertfordshire HP2 4TP, UK
E-mail: info@genstat.co.uk
Website: http://www.genstat.co.uk/

First published 1996, as *GenStat for Windows: an Introductory Course*
This edition published 2020, for Genstat *for Windows* 21st Edition

Contents

Introduction

Genstat is a comprehensive statistical system that allows you to summarize, display and analyse data. The use of the computer for data analysis can save a great deal of time and trouble, but telling a computer what to do can be a troublesome business in itself. General-purpose computing languages, such as Fortran or C++, are designed to deal with the details of arithmetic and communication between a person and a computer; but quite ordinary methods of analysis need long programs. Specialist statistical packages are designed to provide an easy-to-use environment, where only a few instructions or selections from a menu are needed to do a standard analysis; but for something different from the standard, packages are difficult or even impossible to change.

The Windows™ implementation of Genstat gives you the best of both worlds: the flexibility of a programming language with the simplicity of operation of a menu-driven package. It provides this through a standard Windows™ interface, with multiple windows and menus for standard analyses. The menus generate commands automatically to carry out the actions you choose, using Genstat's high-level statistical programming language. However, the command language is also available for you to construct your own analyses simply and concisely, when you want something new or non-standard.

Here are some of the things that Genstat can do:

- manage data, entered by Genstat's own spreadsheet or imported from existing computer files;
- illustrate data with graphics such as histograms, boxplots, scatter plots, line graphs, trellis plots, contour and 3-dimensional surface plots;
- summarize and compare data with tabular reports, fitted distributions, and standard tests, such as t-tests, χ^2-tests and various non parametric tests;
- transform data using a general calculation facility with a wide range of mathematical and statistical functions;
- model relationships between variables by linear or nonlinear regression, generalized linear models, generalized additive models, generalized linear mixed models or hierarchical generalized linear models;
- analyse experiments, ranging from one-way analysis of variance to complex designs with several sources of error variation, using a balanced-ANOVA or a REML approach (including the modelling of correlation structures);
- design investigations deciding on the sample size, or numbers of replicates, required to detect the anticipated treatment effects;
- identify patterns in data by means of multivariate techniques such as canonical variates analysis, principal components analysis, principal coordinates analysis, correspondence analysis, partial least squares, classification trees and cluster analysis;
- analyse results from stratified or from unstructured surveys;
- plot control charts, print Pareto tables and calculate capability statistics;
- analyse time series, using Box-Jenkins models or spectral analysis;
- analyse repeated measurements, by analysis of variance, or using ante dependence structure, or by modelling the correlation over time;
- analyse spatial patterns, using Kriging or spatial point processes.

These techniques are useful in agriculture, ecology, genetics, medical research, and other areas of biology, as well as in industrial research and quality control, and economic and

social surveys; in fact in any field of research, business, government or education where statistics are used.

The version of Genstat described here is the Twenty first Edition for PCs under Microsoft Windows. Its menus are based on an underlying command language, Genstat Release 21, which is available for you to use for non-standard analyses. In this book, we introduce the command language only briefly, at the end of each chapter. To learn more you could read the *Introduction to the Genstat Command Language*. Alternatively, there is a full, formal description in the *Guide to the Genstat Command Language: Part 1 Syntax and Data Management*. This language is common to all implementations of Genstat, including those on workstations and mainframes. The second part of the Guide (*Part 2 Statistics*) gives a comprehensive account of the statistical content of Genstat, reviewing the underlying methodology, explaining the output, and describing the relevant Genstat commands. There are also several specialized Guides, e.g. for *ANOVA and Design*, for *Regression, Nonlinear and Generalized Linear Models*, for *REML* (analysis of linear mixed models), for *Multivariate Analysis*, for *QTL Analysis*, for *Microarrays*, for *Graphics* and for the *Genstat Spreadsheet*. These books are all accessible, in PDF format, from the Help menu (Figure 1.6).

Note (in this section and later in this book): Microsoft, Windows, Windows7, Windows 8, Windows 10, Excel, Explorer, Word and Access are trademarks or registered trademarks of Microsoft Corporation.

1 Getting started

1.1 Using menus

Figure 1.1

You start Genstat within Windows on a PC by selecting the Genstat icon, shown in Figure 1.1, from the Programs Menu or double-clicking on the icon if it appears on the desktop. This runs two processes on the PC. The first, known as the Genstat *Client*, controls the Windows interface for Genstat. It collects information from you, and sends it to the Genstat *Server*, which runs in the background and performs the calculations. Figure 1.2 shows the screen that appears under Windows 8.1 once Genstat has started (Windows 10, Windows 8 and earlier versions are similar). You can see the icon of the Client in the main part of the task bar, and the icon of the Server in the tray on the right-hand side.

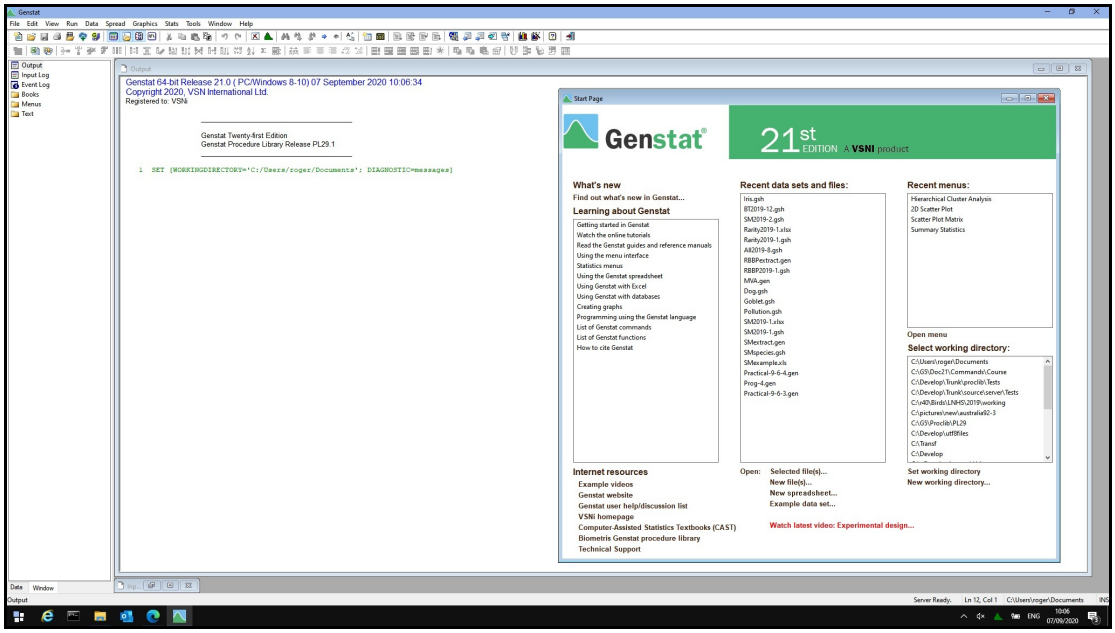


Figure 1.2

Figure 1.2 shows that the Client provides a standard Windows interface, with a menu bar and tool bars at the top, and a status bar at the bottom. There are two tool bars. The lower tool bar is for the Genstat spreadsheet, which is described in more detail in the *Guide to the Genstat Spreadsheet*. There may be several sub-windows. Initially Genstat displays a Start Page, which allows you to perform actions like opening a data file or accessing some helpful documentation. You can close the page by clicking on the red cross at the top right-hand corner, in the usual way. If you do not want to see this menu in future runs of Genstat, you can uncheck the box Show start page box on the General tab of the Options menu; see Figure 1.38, at the end of this chapter.

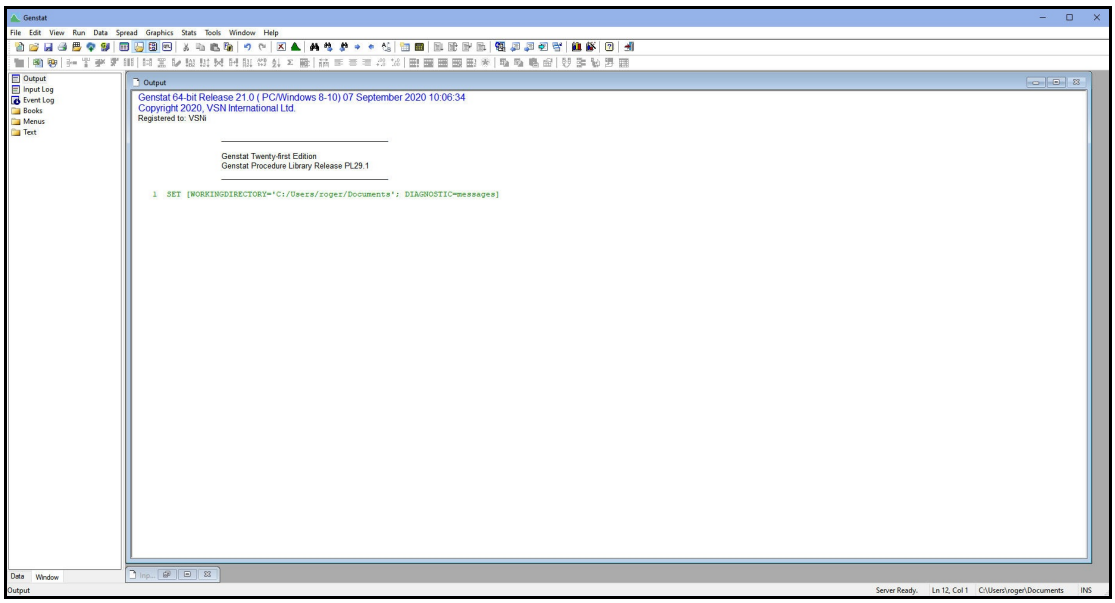


Figure 1.3

Otherwise, Genstat displays a single window, called Output, as shown in Figure 1.3. This initially contains information about the version of Genstat. Later it will contain output from the operations that we perform. The title bar of the window is highlighted to show that it is the current window, and the cursor can be seen blinking inside it. The left-hand section of the status bar at the bottom also shows which window is current; the other sections show the status of the Genstat Server, the position of the cursor in the current window, the working directory, and whether the current window is in insert or overwrite mode.

The Output window can display output in either rich-text (RTF) or plain-text styles. The rich-text style, which is compatible with word-processing systems like Microsoft Word, is shown in Figure 1.3. Titles and captions appear in large, bold or coloured fonts, and columns of output are separated from each other by tab characters. The formatting is thus preserved if you copy or paste into word-processed documents. Alternatively, some users prefer the simplicity of the plain-text style. Genstat then displays the output in a font such as Courier, where all the characters have equal widths, and uses space characters between columns. Section 1.7 explains how to switch between the two styles, and how to change the fonts used for the output. We will use the plain-text style when we show output later in this *Introduction*, to ensure that it is distinguished more clearly from the text of the *Introduction* itself.

In Figure 1.3, the panel on the left-hand side of the main Genstat window is displaying the Window Navigator, which provides an easy way of opening any of the currently available windows. Alternatively, you can click on the Data tab to display the Data View pane, as shown in Figure 1.4. This shows the data currently available inside Genstat. Later in this chapter we will examine the Input Log, which records the Genstat commands that the Client sends to the Server to carry out your requests. This is currently iconized below the Output window.

You can change the appearance of the Genstat window by the usual Windows methods. If you click on the box-shaped icon in the top right-hand corner, then as each window becomes the current window, it will always fill the whole screen(see Figure 1.4). Alternatively, you might resize the windows so that you can view several at once: for

example you might arrange the Input Log below the Output window. Another possibility is to click on the Attach to Frame line in the Window menu (see Figure 1.35 in Section 1.7). The current window is then resized so that it fills the whole frame, but other windows retain their existing sizes and are superimposed over it when they become the current window.

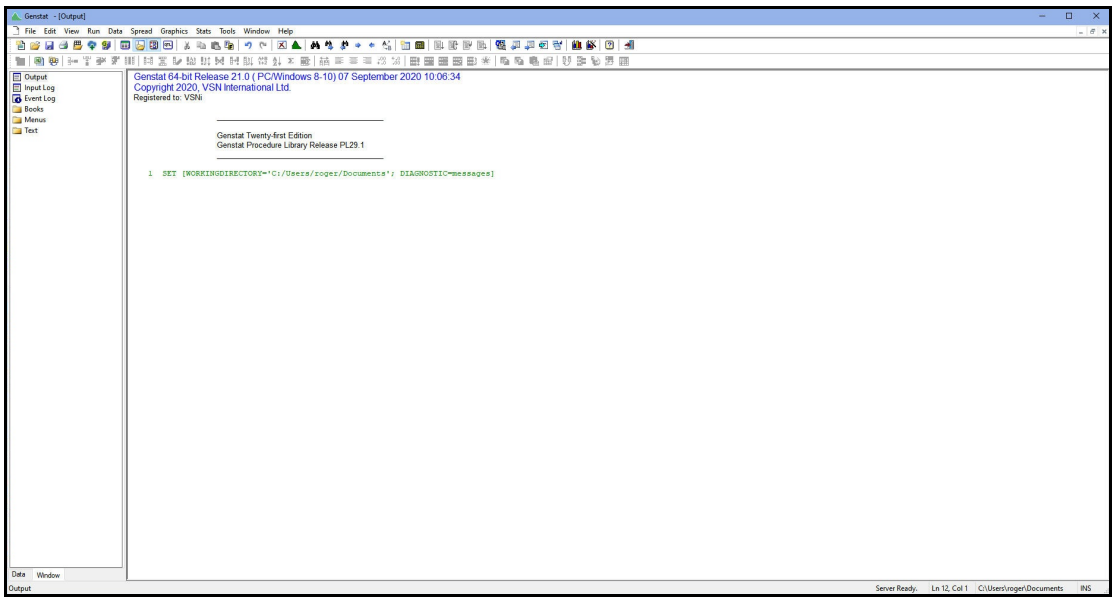


Figure 1.4

You can close either the Data View or the Window Navigator panes by making a right-mouse click on its tab (Data or Window), and clicking Hide on the resulting menu as shown in Figure 1.5. Alternatively, you can uncheck their entries in the View menu on the menu bar (Figure 1.32).

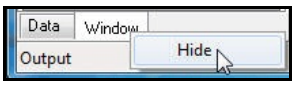


Figure 1.5

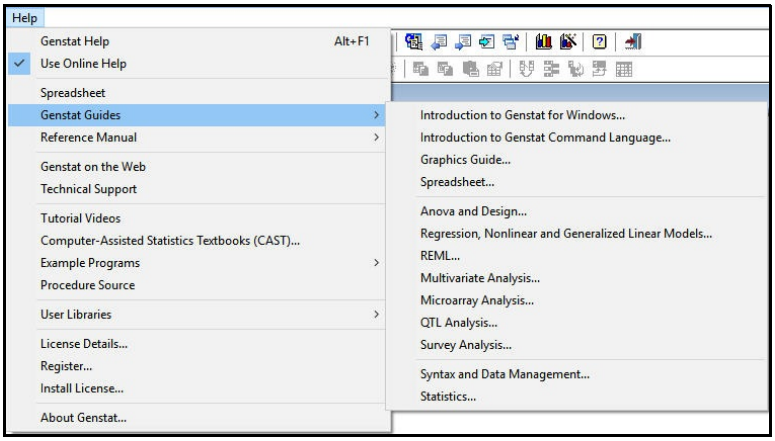


Figure 1.6

Many of the menus provided from the menu bar are standard for Windows applications. Section 1.7 describes these briefly, and the rest of this book introduces many of the other, Genstat-specific menus for carrying out statistical analysis and presentation. To start with, though, it would be useful to become familiar with the on-line help provided by the Help menu – the right-most pull-down menu on the menu bar, shown in Figure 1.6. This provides access to several sources of information (including this book, in PDF format, by clicking on Introduction to Genstat for Windows).

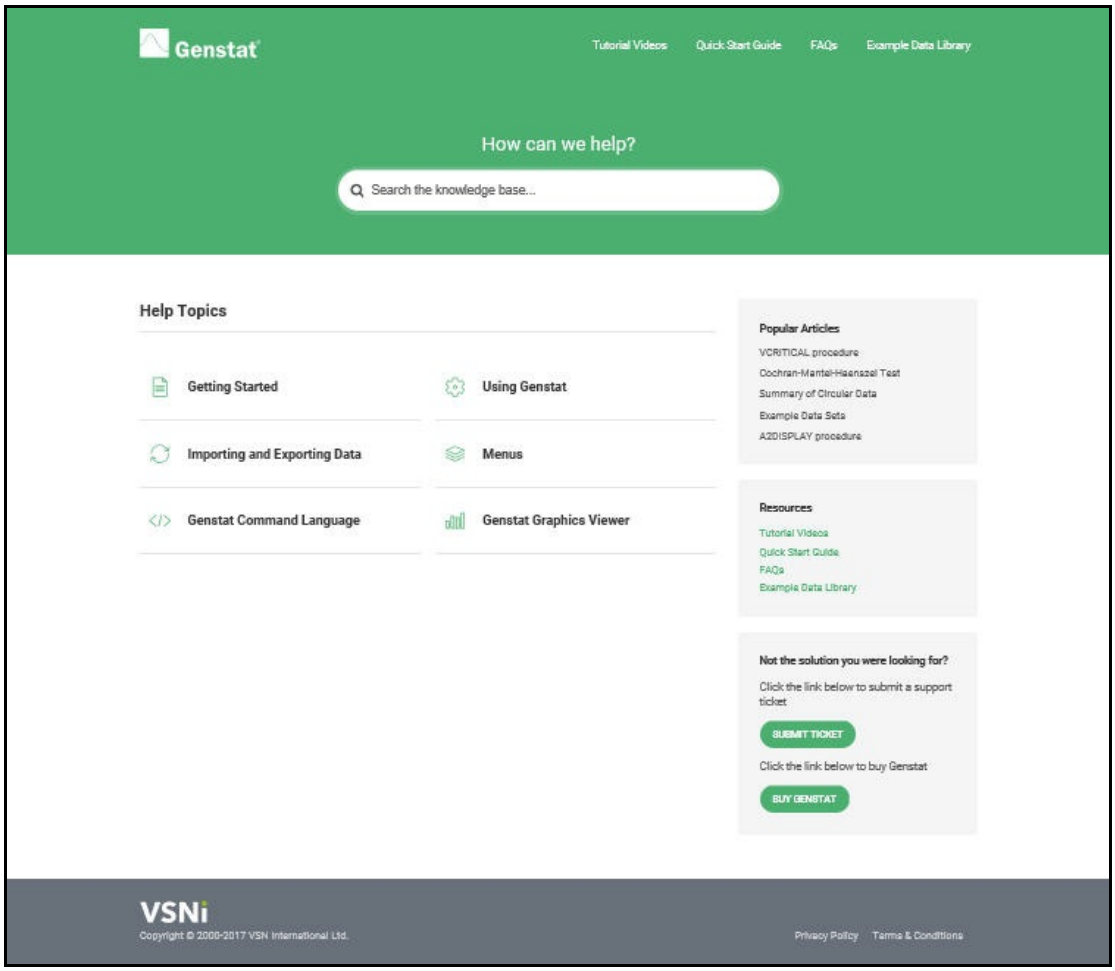


Figure 1.7

If you click on Genstat Help, Genstat opens the VSNi Knowledge Base in your web browser, at the Genstat Version 21 main page, as shown in Figure 1.7. You can use the contents menu on the left-hand side to access the various sections:

Getting Started	explains how to use the Knowledge Base, and shows a simple example;
Using Genstat	describes the Genstat interface;
Menus	provides help information on the menus;
Importing and Exporting Data	explains how to get data into Genstat, and how to save results;
Genstat Language Reference	gives details about the Genstat command language, and the methods that it provides;
Genstat Graphics	describes the sub-system that plots graphs for Genstat; and
Genstat License Management	provides information about the Genstat license-key system.

Alternatively, you can click on the Index button, to see an alphabetical list of the available topics.

There are also interactive tutorials obtained by clicking the eighth line (Tutorial Videos) of the Help menu, which enable you to learn about Genstat by viewing videos.

The first task when you start to use Genstat is usually to access your data. All the data files used in this *Introduction* are stored in a directory (or folder) alongside the folder that contains the Genstat executable program. The menu that opens files always starts in your *working directory*. To make the `Data` folder your working directory, click on Tools on the menu bar and then on the Working Directory line, as shown in Figure 1.8.

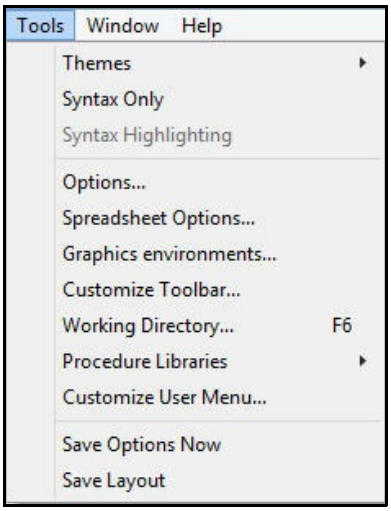


Figure 1.8

This loads the Working Directory menu show in Figure 1.9. Click on the `Add Data Folder` button and then on OK to make the folder your working directory.

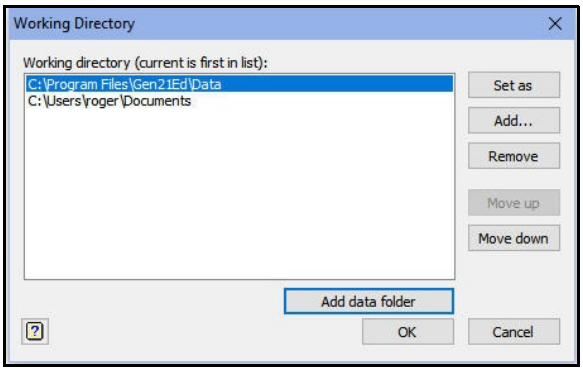


Figure 1.9

In later chapters we show several ways of entering data directly on the screen or importing it from various types of file; but here we shall use the easiest way, importing a set of data that was stored during a previous Genstat session. Clicking Data on the menu bar pulls down a menu with several options (Figure 1.10), selecting Load and then clicking on Data File brings up the Open File menu, which allows you to select a file containing some previously stored data.

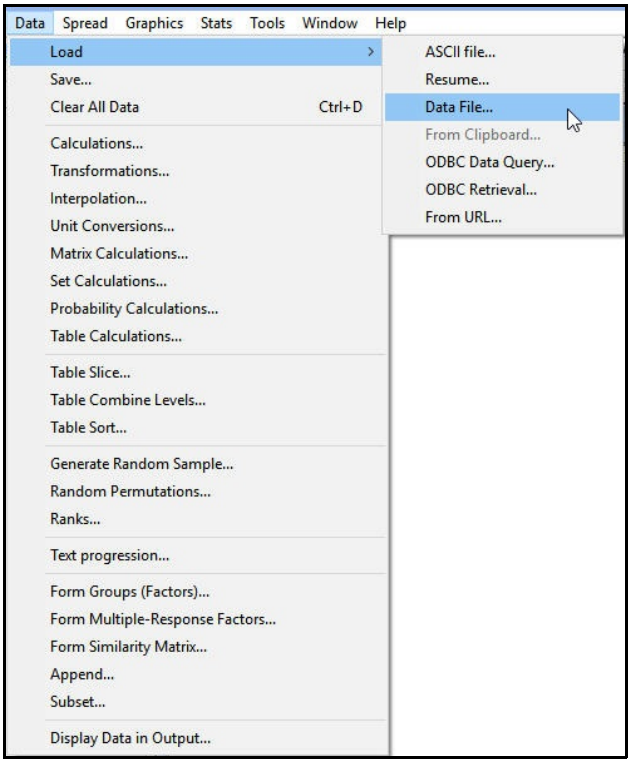


Figure 1.10

The menu now opens in the `Data` folder as your working directory. Notice that the menu provides the standard Windows tools for moving from one folder to another. If you do that, you may then want to click on the Working Directory button, and then Set as in the resulting drop-down list, to request that the folder that you have reached becomes the new working directory. The Go to item in the list returns you to the working directory, and the Select item opens a menu where you can select any of the recent working directories.

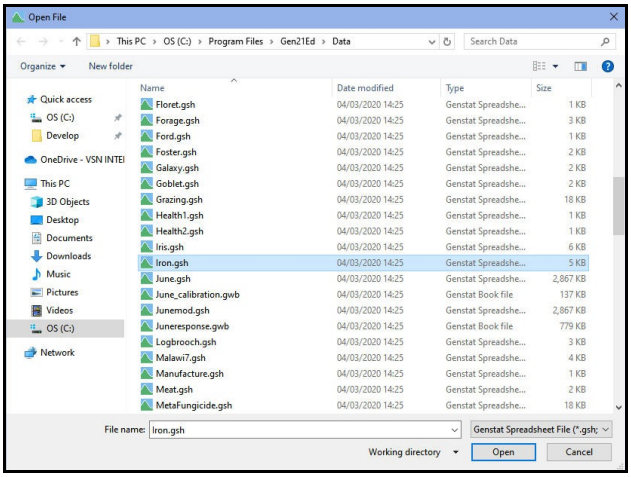


Figure 1.11

In Figure 1.11, we have selected Genstat Spreadsheet Files (*.gsh, *.gwb) from the drop-down list in the File name menu, and then clicked on the file `Iron.gsh`. You can then load the data by clicking on Open. Genstat loads the data from the file and displays a report in the Output window, shown below.

Data imported from Genstat Spreadsheet: C:\Program Files\Gen21Ed\Data\Iron.gsh
on: 7-Sep-2020 10:24:26

Identifier	Values	Missing	Levels
sample	136	0	12

Identifier	Values	Missing	Levels		
site	136	0	6		
Identifier	Minimum	Mean	Maximum	Values	Missing
FE	200.6	246.6	308.2	136	0
Identifier	Minimum	Mean	Maximum	Values	Missing
weight	11.36	12.42	13.00	136	0

There are four sets of information, or *data structures*, in this file. Two of these simply contain a series of numbers: the structure called `FE` stores the measured parts per million (ppm) of iron, and `weight` stores the weights of soil that were analysed. The values of the structures correspond to the 136 soil samples that were analysed in the whole study. The other two structures categorize these samples: each value of `site` records the code number of the laboratory that carried out the analysis, and `sample` contains the number (from 1 to 12) of the originating soil sample that was given to the laboratory to analyse. Notice that the summaries of the categorical structures are different to those of the purely numerical columns. Depending on how your copy of Genstat has been set up, the Output window may also contain a listing of the commands that Genstat has carried out in order to load the data. If this occurs, you may want to use the Options menu, shown in Figure 1.37 (in Section 1.7), to stop the commands being echoed.

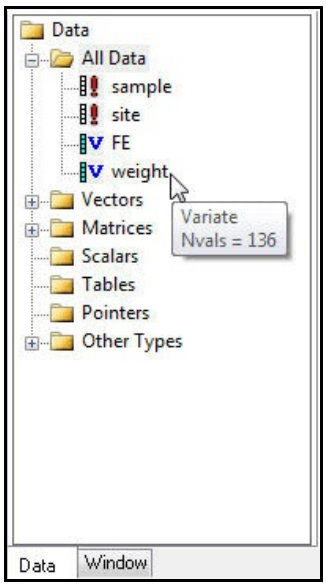


Figure 1.12

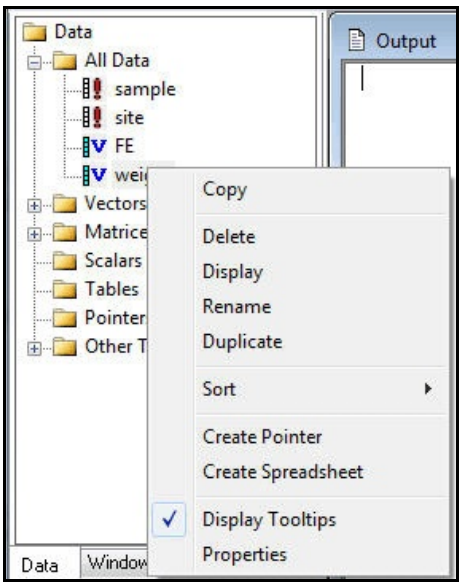


Figure 1.13

The data structures can also be seen in the Data View pane. The tree allows you to select the types of data structure that you want to list on the right-hand side: in Figure 1.12, we have opened All Data.

The structures that store lists of numbers in Genstat are known as *variates* and identified by a blue "v", whereas the categorical structures are called *factors* and identified by the red exclamation mark. Genstat provides a range of data structures that are convenient for different types of data, but these two are the most common. Notice

that, as you rest the mouse on the name of a structure, a small window appears with information about its attributes. These *tool tips* are controlled by the right-mouse menu (Figure 1.13), obtained by making a right-mouse click on any of the structures. In addition, the menu also allows you, for example, to delete, rename or display (i.e. print in the Output window) structures. You can also drag and drop structures from Data View onto most of Genstat's other menus.

Clicking Stats on the menu bar allows you to select any of Genstat's statistical menus. In Figure 1.14, we have selected the Summary Statistics sub-option of the Summary Statistics option, which opens the Summary Statistics menu shown in Figure 1.15. This allows you to display summary statistics describing the contents of a variate, and also to produce some useful graphs. First we have used the -> button to put the name of the variate Fe into the Variates box. (Alternatively, you can double-click on the required variate or variates, but the button allows you to transfer several variates selected using the standard Windows techniques: clicking the mouse with the Ctrl key depressed to add or remove a variate from the selection, or clicking the mouse with the shift key depress to add a contiguous list.). We have then moved the cursor to the By Groups box by pressing the Tab key or clicking on that box, and selected the factor site from the new entries shown in Available data. We have also checked the Boxplot box.

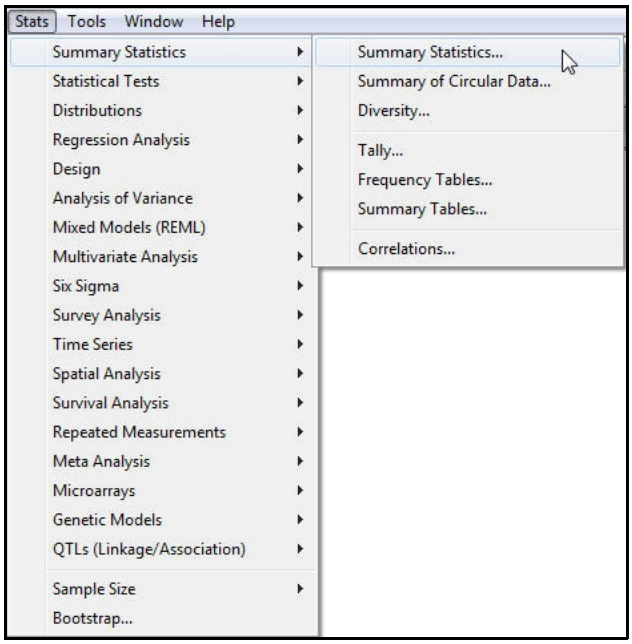


Figure 1.14

This menu, like many of the other Genstat menus that you will see later in this *Introduction*, has four standard icons in the bottom left-hand corner. Working from the left: the Pin button controls whether the menu closes or remains open after it has been run; the Restore button (with the curved-arrow symbol) can be used to restore names into the edit fields and recover default settings; the Clear button (with the red cross, or "rub-out", symbol) clears any menu settings; and the Help button (with the question mark) gives information about how to use the menu. In Figure 1.15, the Pin

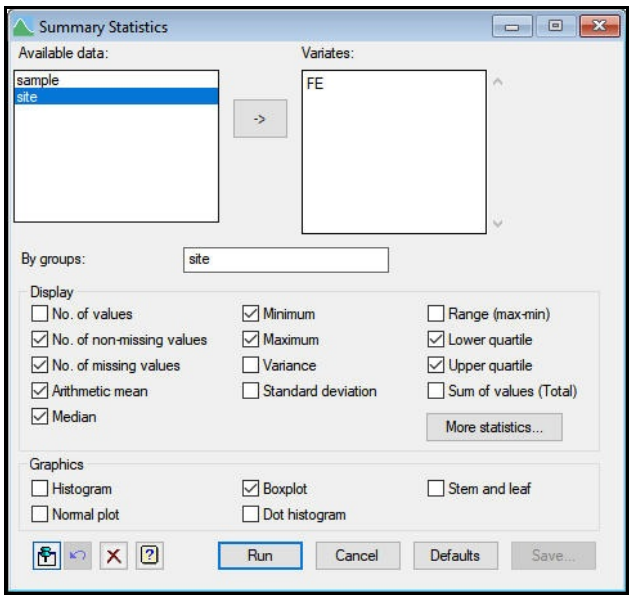


Figure 1.15

button is in the "in" position, with the pin vertical. So the menu will stay open after it has been run.

When you click on the Run button, Genstat prints summary statistics for each site, in turn, as shown below. It also opens a Graphics window and draws six *boxplots* in it, one for each of the laboratories (or sites). This process may take a little more time than some of the operations done earlier, because Genstat has to start its Graphics viewer running first. While this is going on, the Status bar will display the message Processing summary of variates in place of Summary Statistics menu, to let you know that the computations are taking place.

Summary statistics for FE: site 1

Number of observations =	24
Number of missing values =	0
Mean =	289.6
Median =	289.1
Minimum =	269.5
Maximum =	308.2
Lower quartile =	282.1
Upper quartile =	295.6

Summary statistics for FE: site 2

Number of observations =	22
Number of missing values =	0
Mean =	274.2
Median =	273.2
Minimum =	262.6
Maximum =	283.1
Lower quartile =	270
Upper quartile =	280.1

Summary statistics for FE: site 3

Number of observations =	24
Number of missing values =	0
Mean =	216.3
Median =	212.8
Minimum =	200.6
Maximum =	252.7
Lower quartile =	208.7
Upper quartile =	218.4

Summary statistics for FE: site 4

Number of observations =	18
Number of missing values =	0
Mean =	239.5
Median =	238.1
Minimum =	232.5

Maximum =	255.7
Lower quartile =	236.5
Upper quartile =	239.4

Summary statistics for FE: site 5

Number of observations =	24
Number of missing values =	0
Mean =	234.9
Median =	234.6
Minimum =	222.7
Maximum =	251.6
Lower quartile =	230.2
Upper quartile =	237.1

Summary statistics for FE: site 6

Number of observations =	24
Number of missing values =	0
Mean =	225.5
Median =	224.2
Minimum =	215.3
Maximum =	238.6
Lower quartile =	221.8
Upper quartile =	229.1

The output is labelled using standard statistical terminology but, if you are unsure about any of the words or phrases, you can use Genstat's context-sensitive help. Put the cursor into the word of interest, or into the first word of the phrase of interest, and press the F1 key. For example, if you put the cursor into the word "quartile" and press F1, Genstat opens its help system at the Quartile topic, which contains the information: "The lower quartile is the value *l* such that 25% of a sample are less than *l*. Similarly, the upper quartile is the value *u* such that 25% of a sample are greater than *u*."

Sometimes there is more than one potentially relevant topic. Genstat then provides a menu so that you can select the one that seems most appropriate. The menu for the word "median" is shown in Figure 1.16. Selecting Median (explanation from glossary) produces the definition: "Median is the value that divides a sample into two equally sized groups.".

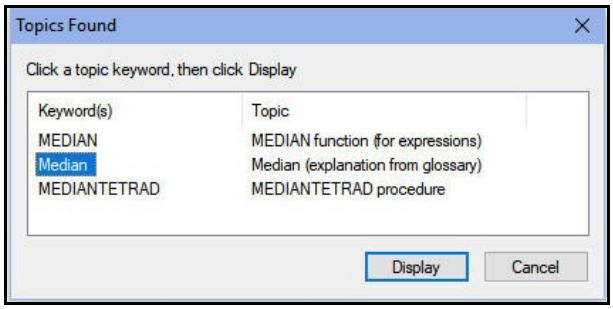


Figure 1.16

Figure 1.17 shows the graph, displayed in a separate window by Genstat's Graphics Viewer. (If you look on the task bar you will see that this has its own icon there.) You can zoom the display using the slider on the toolbar, or by holding the left mouse button and moving the mouse up and down. If your mouse has a centre button, you can move the display within the window by moving the mouse with that button held down. Alternatively, you can use the scroll bars at the bottom and on the right-hand side of the screen.

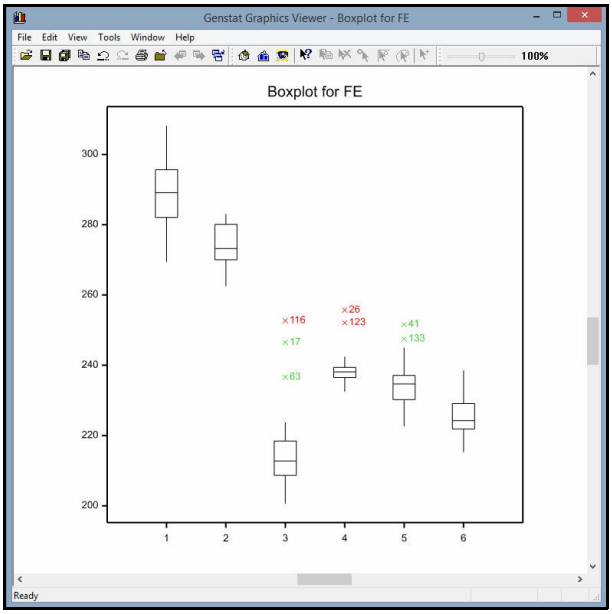


Figure 1.17

These are *schematic* boxplots. Each one has a central box spanning the inter-quartile range of the data from a laboratory (so that 50% of the observations lie inside the box) with a horizontal bar drawn across the box at the median. There are whiskers extending from each box to the most extreme data values within inner *fences* that are defined to be at a distance of 1.5 times the interquartile range beyond the quartiles, or at the furthest data value if that is smaller. Points that lie beyond the whisker are regarded as *outliers*, and are plotted as crosses with labels identifying their unit number. Points that lie beyond outer fences, defined to be at a distance of three times the interquartile range beyond the quartiles, are regarded as *far outliers* and plotted in red. Ordinary outliers are plotted in green. The display shows that Laboratories 1 and 2 are producing consistently higher results than the rest, and Laboratory 3's results are generally lower.

To return to the main Genstat screen, you can click the Genstat icon on the task bar. Then click Cancel in the Summary of Variates menu to remove this menu, if you no longer need it.

1.2 Practical

Select the Data File sub-option of the Load option of the Data menu. Move from the directory (i.e. folder) containing the Genstat executable program to the `Data` directory and set this as the working directory.

Load the file `Sales.gsh` and open the Data View pane to see what data structures it contains.

Use the Summary of Variates menu to display the minimum, maximum and mean sales, and to plot a boxplot of the sales in each town.

Print the skewness and kurtosis of the sales figures, and use the context-sensitive help to obtain some information about what these represent.

1.3 Giving commands

As well as using menus, or instead if you prefer, you can tell Genstat what to do by giving it *commands*. In fact, the menus themselves work by constructing commands automatically and sending them to the Genstat Server. You can see these commands in the Input Log; by default you cannot edit within this window (it is "read only"), but this can be changed using the Options menu (see Figure 1.36). The contents of this window after drawing the boxplot are in Figure 1.18.

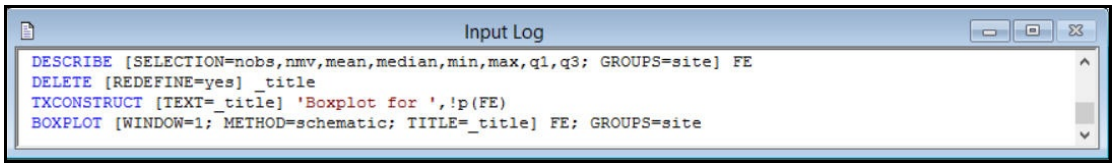


Figure 1.18

There are four commands here: to print the summary statistics (DESCRIBE), to delete the temporary structure that will be used to contain the title for the boxplot (DELETE), to form that title (TXCONSTRUCT), and to draw the boxplot (BOXPLOT). Notice that the name of the temporary structure (_title) begins with the underscore character (_). Genstat always does this with temporary structures, to help distinguish them from your own data structures. Previous commands to set the working directory and to load data into Genstat from the file have scrolled up above this narrow window.

We introduce the Genstat command language only briefly here. To learn more you should read the *Introduction to the Genstat Command Language,* which is a companion to this Introduction.

All Genstat commands have a common form of *syntax*: in other words, there are some general rules that apply to all the commands that you give. We introduce the basic syntax here. Further details are given in the *Introduction to the Genstat Command Language*. There are two types of commands: *directives* are the basic commands of the Genstat language while *procedures* are extensions of the language, using programs written in the Genstat language itself. However, both obey identical rules, so you do not need to know which you are using.

Genstat can use different colours to identify the various components of the command. Here, for example, the names of the commands are in blue. This *syntax highlighting* of the current window can be controlled by checking or unchecking the Syntax Highlighting line of the Tools menu (see Figure 1.35).

We introduce the rules in the context of the directive called PRINT. This displays data, allowing you to inspect individual values. Note that this command does not send data or results to a printer: to do that you can select the Print option from the File menu in the menu bar. There is also a menu to display data values, opened by clicking on the Display Data in Output option of the Data menu on the menu bar (Figure 1.10).

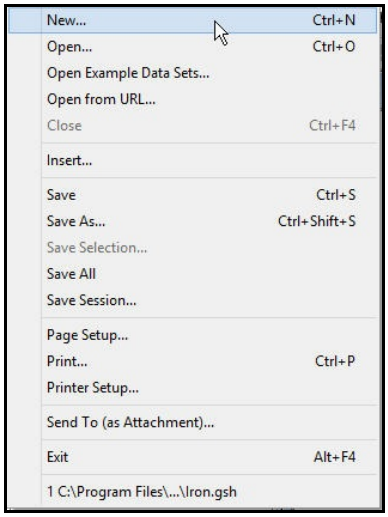


Figure 1.19

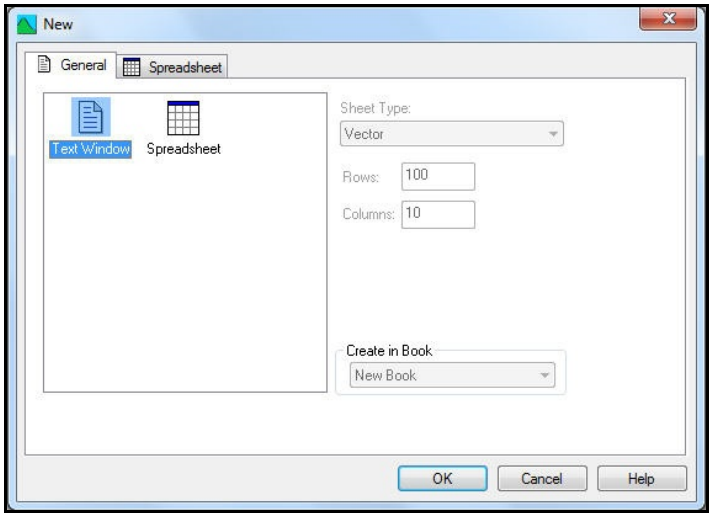


Figure 1.20

To give commands directly, it is best to open a new window in which to construct them. This is done by clicking on File in the menu bar and selecting New, as shown in Figure 1.19. This generates the menu shown in Figure 1.20, allowing you to choose what type of new window you want. Selecting Text Window and clicking on OK gives you a new, empty, window which will become the current window. You can type, for example, the simple command

```
PRINT 1
```

to display a single set of data: the number 1.

When constructing commands in a window, you can use the usual keys for typing and deleting characters, and moving about the window. You can also switch between Insert and Overwrite mode by pressing the Insert key, and the Status bar will display, with Ins or Ovr, which mode you are in at any time.

This is a trivial exercise, of course, but it serves to show how commands work. To get Genstat to execute this command, leave the cursor at the end of the line (that is, just after the 1) and select the Run menu from the menu bar. Select Submit Line, as shown in Figure 1.21, and the command will be executed. The resulting display is put in the Output window, as shown in Figure 1.22.

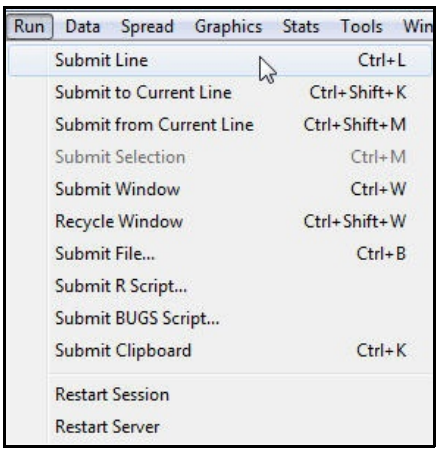


Figure 1.21

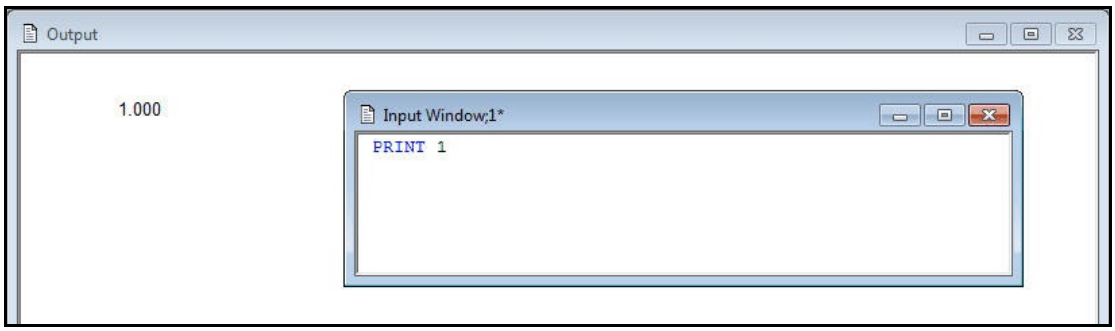


Figure 1.22

This is clearly not a very useful operation, because you already know what the set of data is, and because it consists only of a single number; however, this will quickly be generalized. In the meantime, you can see that the directive name, `PRINT`, is like a command verb which instructs Genstat to do something, and the number `1` is like the object of the command. All directives, and procedures, work like this, though not all directive names are actually verbs in the English language. The object is called the *primary parameter* of the command.

The `PRINT` directive, like all others, works with sets of data. You can make it work with several sets of data at once by giving a *list*; for example, the command

```
PRINT 1,2
```

has two sets, each containing one number, as shown in Figure 1.23.

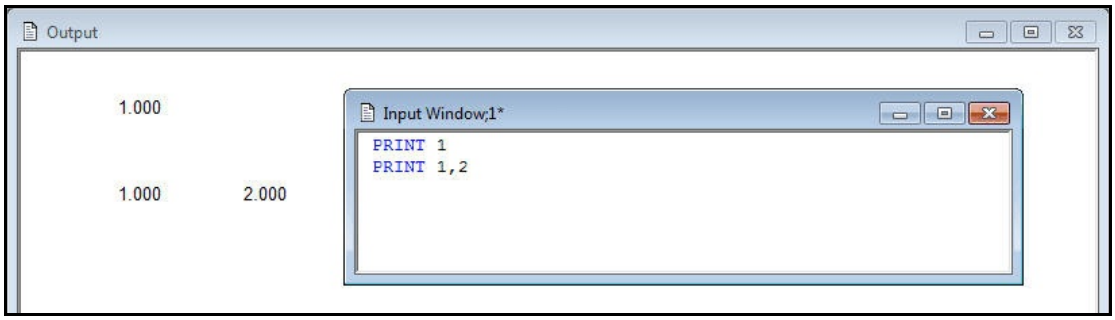


Figure 1.23

In Genstat, lists are always constructed using commas. You must not use just spaces; for example, the command

```
PRINT 1 2
```

would be faulted, because the space may be an accident, and you may have meant

```
PRINT 12
```

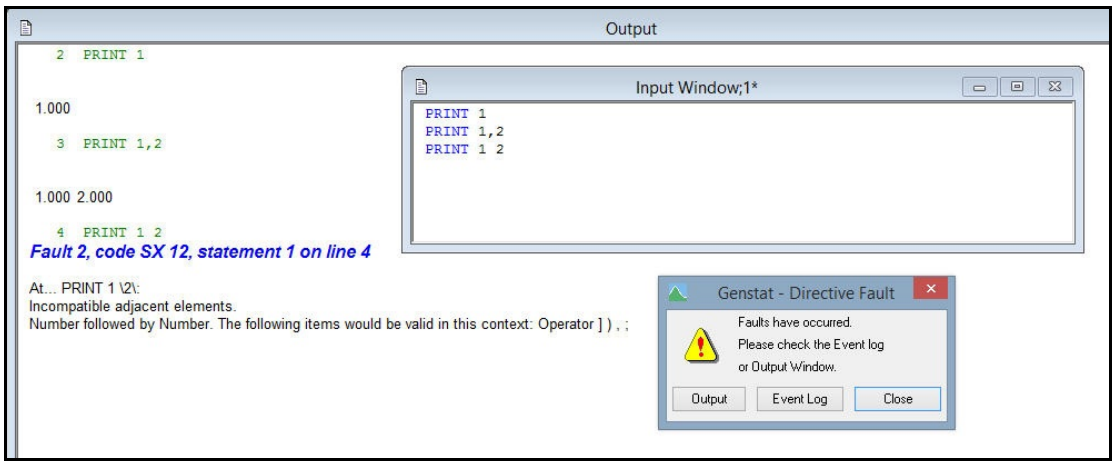


Figure 1.24

If you make a mistake like this, Genstat puts a brief explanatory message in the Output window, and records the fact that a fault has occurred in the a separate window called the Event Log (see Figure 4.9 in Section 4.1 for an example). The message is shown in Figure 1.24. You can click on the Output button to go to the fault in the Output window, or on the Event Log button to open the Event Log.

You can, though, use spaces as well as commas if you want, so the following command is acceptable:

```
PRINT 1 ,    2
```

You will have noticed that `PRINT` commands lay out the data in a tabular form, choosing an appropriate number of decimal places for numbers. By default, a single number is displayed with four significant digits. Also, sets of data with compatible shape are laid out in *parallel*: that is, side-by-side. If you don't want this default display, there are a range of *options* to modify it. For example, the command

```
PRINT [SERIAL=yes] 1,2
```

one number by itself, and then the other, as shown in Figure 1.25.

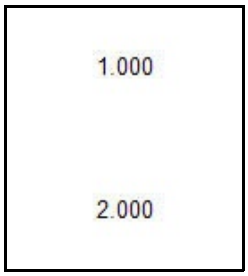


Figure 1.25

Most Genstat directives and procedures have options like this to control the way in which the operations are done. They must always be given in square brackets following the directive or procedure name and preceding the parameters, if any. Options have the form *name=setting*, where here the name is `SERIAL` and the setting is `yes`. Settings can be words, as here, or numbers. If you set several options, you must separate them with a semi-colon, as in

```
PRINT [SERIAL=yes; INDENTATION=10] 1,2
```

This command would indent the output by 10 characters, so that if you arrange to send the display to a printer, you could rely on having a clear margin on the paper, perhaps for binding.

The `CHANNEL` option of `PRINT` was used in the Input Log, to put the output into the Genstat *text* structure `_tmptext`.

Most Genstat directives and procedures also have *auxiliary parameters* which control the way the command works. For example, the command

```
PRINT 1,2; DECIMALS=0,1
```

gives the output shown in Figure 1.26.

Figure 1.26

The effect of the DECIMALS parameter is to specify how many decimal places to display for each set of data. The essential difference between an option and an auxiliary parameter is that an option specifies a modification for all the sets of data in the command, while an auxiliary parameter specifies a modification that may be different for each of the sets of data in turn. The setting of the DECIMALS parameter above, 0,1, is matched item by item with the setting of the primary parameter, 1,2. This distinction applies to all Genstat commands.

The setting of an auxiliary parameter is otherwise like that of an option, with the form *name=setting*, and the semi-colon separator is needed between successive parameters. The primary parameter itself has a name, except when there are no auxiliary parameters. So you could actually give the command:

```
PRINT STRUCTURE=1,2; DECIMALS=0,1
```

However, if you specify the primary parameter first in a command, its name can always be omitted.

You can abbreviate directive and procedure names to the first four characters; names of options and parameters can also be abbreviated to four characters, and sometimes further. The full abbreviation rules are described in Section 3.3 of the *Introduction to the Genstat Command Language*.

So far, we have used the very simplest sets of data, consisting of a single number each. Most practical work is done with series of numbers, like those in Section 1.2. For example, we can display the values in the variate called FE by simply giving the command:

```
PRINT FE
```

We gave the name FE to this set of data before saving it in the file Iron.gsh: such a name is referred to as an *identifier* in Genstat. You need to give identifiers to data even when using menus, so you should be aware of what Genstat allows. You can use names consisting of up to 32 letters or digits, but they must start with a letter. Case is significant, so the identifier FE is different to fe. We have used capital letters for this identifier but lower case for the others, like sample; however, you may find it easier to stick to all lower-case or all upper-case for your identifiers, at least while you get started with the system.

The PRINT command works on all types of Genstat data structures, so you can probably guess that the following command would display all the data that was loaded in Section 1.1.

```
PRINT sample,site,FE,weight
```

Part of the display is shown in Figure 1.27.

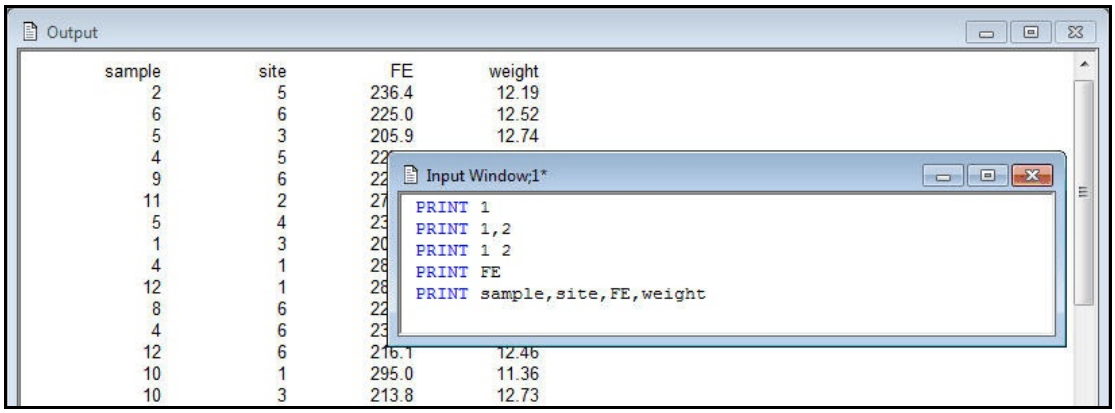


Figure 1.27

Values can be assigned to data structures in many ways; for example, by loading data from a file, by saving the results of an analysis, or by doing calculations (Chapter 2). In Genstat, calculations are specified in *expressions*, and are most often carried out with the `CALCULATE` directive. As a simple example, we can calculate the amount of iron in each tested quantity of soil by multiplying the weight by the concentration of iron:

```
CALCULATE iron = FE * weight
```

The weights of soil were in grams, so the resulting weights of iron will be in micrograms.

This `CALCULATE` command is more powerful than it looks. In fact, 136 separate calculations are done here: Genstat knows to do this because `FE` and `weight` have been defined to be variates with 136 values. So each of the 136 values stored in `FE` is multiplied by the corresponding value in `weight` and the results stored successively in a new variate called `iron`.

1.4 Practical

The price charged for each item sold in the data set `Sales.gsh`, used in Practical 1.2, was 2.99. Calculate the amount received on each day in each of the towns, and use the `PRINT` directive to display day, town and amount received, in columns, in the Output window.

1.5 Working with programs

Instead of making Genstat execute one command at a time, and *interacting* with the results, you can arrange to run several Genstat commands at once. Open an edit window, as before, but construct in it the whole series of commands that you want to execute. If you like, you can store the commands for future reference in a file, by clicking on File in the menu bar, and selecting Save As.

You can tell Genstat to execute a selected set of commands by highlighting them (by clicking at one corner of a block of commands and dragging with the mouse to the opposite corner). If you then choose Submit Selection from the Run menu, the commands are sent to the server, and the results are displayed in the Output Window. Alternatively, you can run all the commands in the window by choosing Submit Window from the Run menu.

There are many example files of Genstat programs which are stored in a folder called

Examples alongside the Data folder. These can be loaded automatically into an input window, and then (if you want) executed, by clicking Help on the menu bar and clicking on the Analysis Programs sub-option of the Example Programs option (see Figure 1.28). This generates the menu shown in Figure 1.29, which enables you to select the example that you want. Here we have selected SORT directive from the Manipulation of data topic.

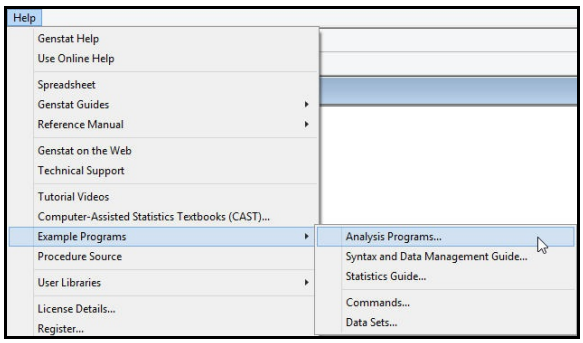


Figure 1.28

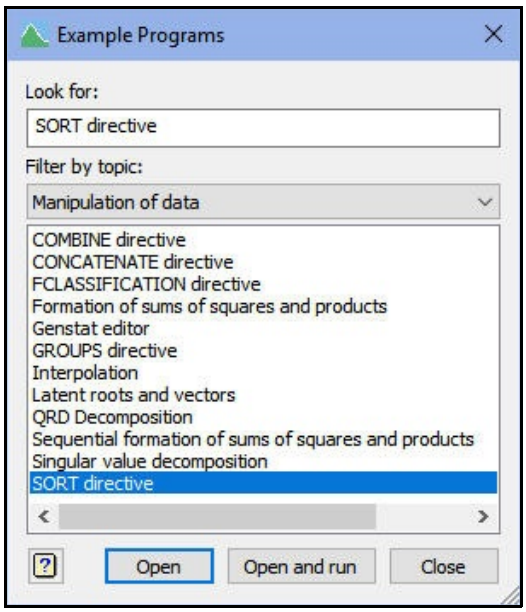


Figure 1.29

Clicking on Open and Run loads the program into a new Input window (Figure 1.30), and runs the program. The resulting output appears in the Output window.

The Tutorial Videos sub-option in Figure 1.28 opens a page in the Knowledge Base with links to videos illustrating the use of menus to perform various types of analysis.The Syntax and Data Management Guide and Statistics Guide sub-options allow you to obtain the examples used in the two Guides that describe the Genstat command language. The Commands sub-option gives examples of individual Genstat commands, and the Data Sets option provides an alternative way of loading the example data sets used in this Introduction and the various Guides.

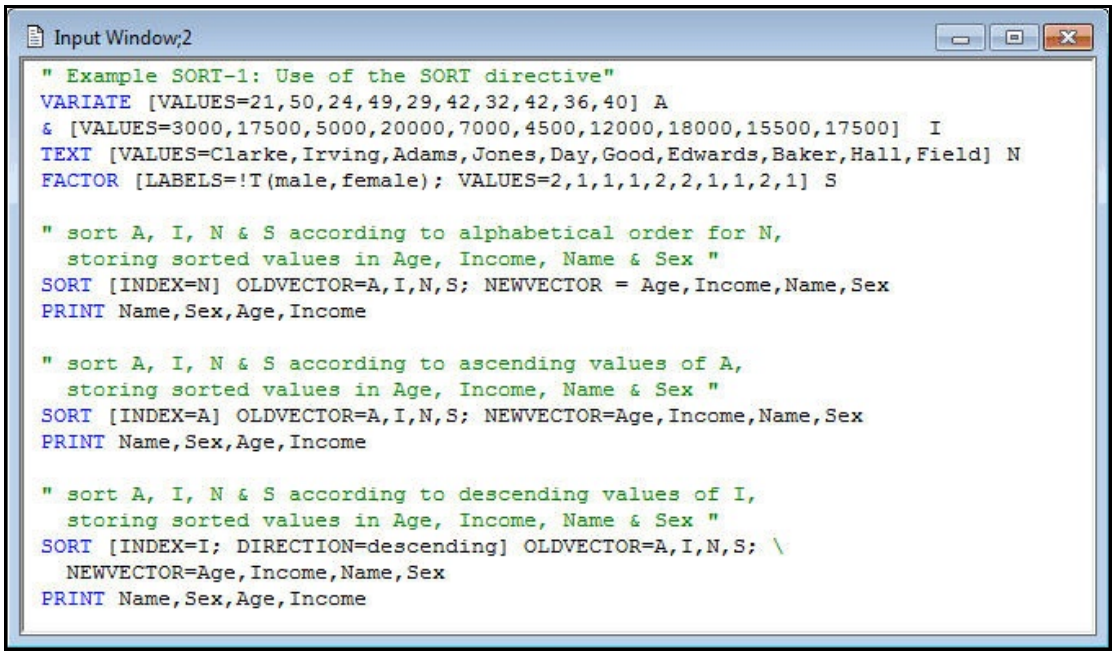


Figure 1.30

1.6 Practical

Click on Help on the menu bar, and then the Syntax and Data Management sub-option of the Example Programs option. Open and run Example 3.2.1a-e `PRINT` directive.

1.7 The Windows interface

The previous sections have introduced many of the features of Genstat's interface, and others will be introduced in later chapters. But we introduce here some general features that are common to many Windows-based programs, which can be very useful when using Genstat in practice, and briefly list the range of statistical techniques that Genstat provides by menus.

The Edit menu, shown in Figure 1.31, allows you to exchange information with other Windows-based programs. For example, the Copy option allows you to copy results from the Output window into a word-processing package. You can select the results to be copied by highlighting them using the mouse, or the Shift key together with the arrow keys, in the Output window. As usual with programs within Windows, the copy is sent to the *Clipboard*, which is a utility provided by the Windows system to help communication between applications.

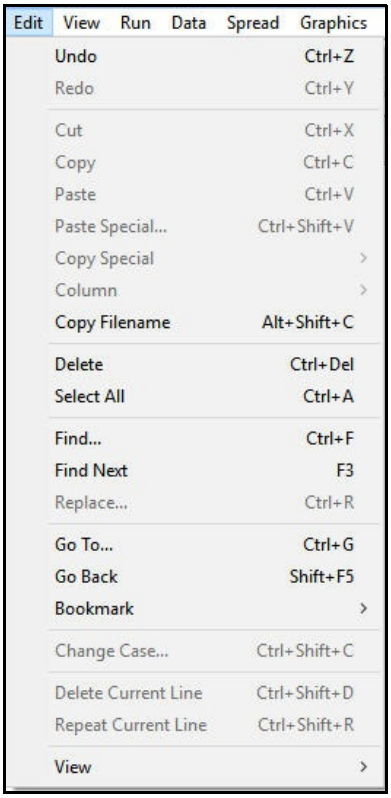


Figure 1.31

Genstat and many other applications provide a Paste option to insert the contents of the clipboard into the current window, at the position of the cursor. In this way, information can be moved within one Genstat window, sent from one Genstat window to another, exported from Genstat to another program, or imported from another program into Genstat.

The Cut option works like Copy, except that a selection must be made, and the information is removed from the current window as well as being copied to the Clipboard. If no selection has been made, the Cut option is displayed in grey rather than black, indicating that it is not yet available; similarly, Paste is not available if there is no material in the Clipboard. The Delete option removes information without putting it in the Clipboard. Any of these operations (except Copy) can be undone by choosing the Undo option.

The Find option allows you to search the current window for a given string, with options to specify case matching and occurrence as a word rather than part of a word. Repeated searches can be done with the Find Next option, and strings can be replaced with the Replace option.

The Go To option allows you to move to a specific line in the current window, if you know its line number (as displayed on the Status bar). The Go Back option is relevant to the Genstat spreadsheet (Chapter 4), enabling you to undo a Go To operation. The Bookmark option allows you to maintain markers in a window, which can be useful if you are generating a long report in the Output window, for example.

The Change Case option can change the letters in a selected string from lower case to capitals, or from capitals to lower case, or into title case.

The View menu is shown in Figure 1.32. The Output option allows you to change the Output window from rich text to plain text or from plain text to rich text. In Figure 1.32, the change will be to plain text.

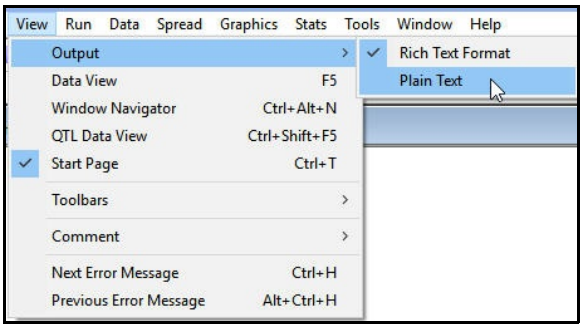


Figure 1.32

Genstat must restart the server to make the change. This will lose all the data currently stored in the server. So Genstat pops up the menu in Figure 1.33 to check whether you really want to do this.

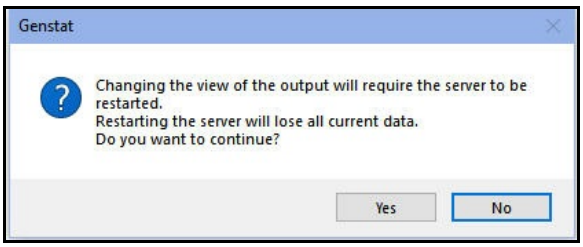


Figure 1.33

The Data View and Window Navigator options allow the Data View and Window Navigator panes to be displayed (Figures 1.2 and 1.3). The Toolbars option controls which toolbars are displayed, and the final options (Next Error Message and Previous Error Message) allow you to move from one error message to another in the Output window.

The Window menu shown in Figure 1.34 allows you to rearrange the sub-windows in the Genstat window, either *cascading* them so that they overlap except for the top left corners, or *tiling* them so that you can see the whole of each window. You can also bring any of the named windows to the top of the display, which is useful if any window has become buried by other windows so that you cannot click the mouse on an exposed part. The Next and Previous options allow you to cycle through all the windows (except graphics), including input windows and dialogue boxes, and the Windows option brings up a list of all windows to choose from, including any open dialogue boxes.

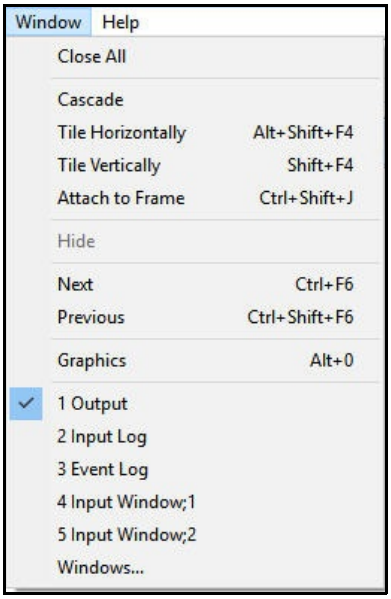


Figure 1.34

Many of these standard Windows options are provided also by buttons on the tool bar, to make it easier to execute them. Genstat provides "tool tips". So, if you leave the mouse pointer on one of the buttons, a small window will appear describing the purpose of the button concerned.

The Tools menu, shown in Figure 1.35, controls various aspects of Genstat, and the environment in which it runs. The Themes option allows you to select, save or load the "theme" to control Genstat's general look-and-feel. There are several pre-defined themes, selectable from the Genstat Theme menu, which is loaded if you click on Select, as in Figure 1.35. You can also save a theme to an external file so that, for example, you can load it on another computer. In this Guide, we are using the "Modern theme".

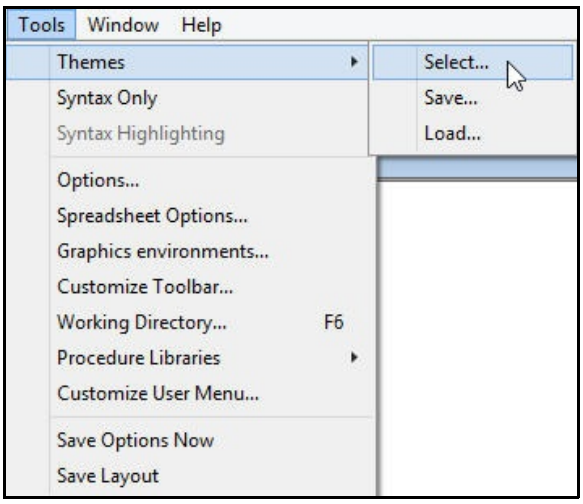


Figure 1.35

The Customize Toolbar option brings up a menu that enables you to choose which buttons appear on the toolbar, and how they are arranged. As we have seen already, in Section 1.1, the Working Directory option provides a method of specifying the working directory, which is where all browse menus first start. The Spreadsheet Options option allows you to control the way in which the spreadsheet operates, and the Procedure Libraries option allows you to connect your own procedure libraries. Clicking on the Syntax Only option requests the client to copy the commands generated from the menus to the Input Log, but not to send them to the server. The Syntax Highlighting option controls the use of different colours to highlight the various elements of the command in any text window (other than the Output Window). The Save Options Now option allows you to carry all the option settings that you have changed using these menus through to your next session with Genstat, while the Save Layout option simply saves the way in which the windows are arranged.

The Options option brings up a tabbed menu with more detailed settings, as shown in Figure 1.36.

For example, the Audit trail tab lets you specify what aspects of your work will be recorded in the Input Log. Here all the commands are being recorded in the Input Log, but they are not being echoed in the Output window, nor is Genstat sending out the special code to start different sections of output on new pages when they are printed.

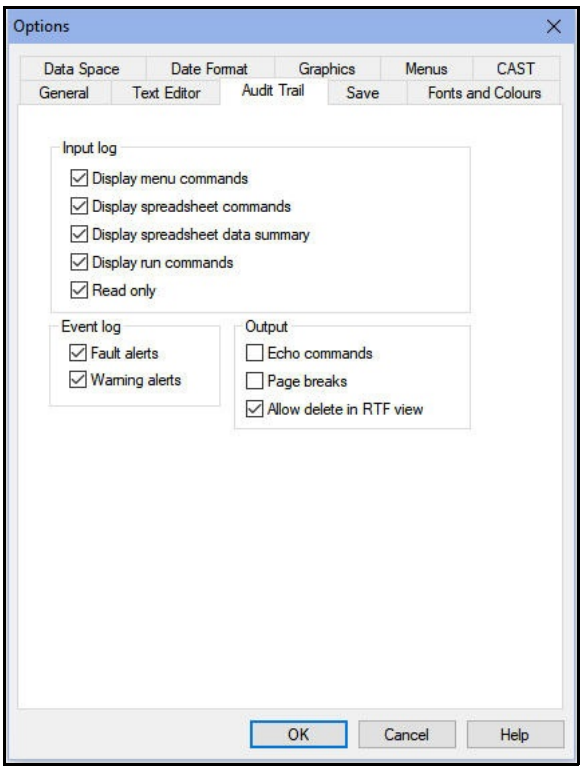


Figure 1.36

The General tab (Figure 1.37) allows you to stop the Start Page appearing when you open Genstat, and to control aspects like the number of files that are saved for the recent file list. It also controls the error reports that Genstat makes. As a general rule, it is best to get faults, warnings and messages. However, you can suppress warnings or warnings and messages if, for example, you are confident that they will be unimportant in your analysis, and want to keep them out of your output.

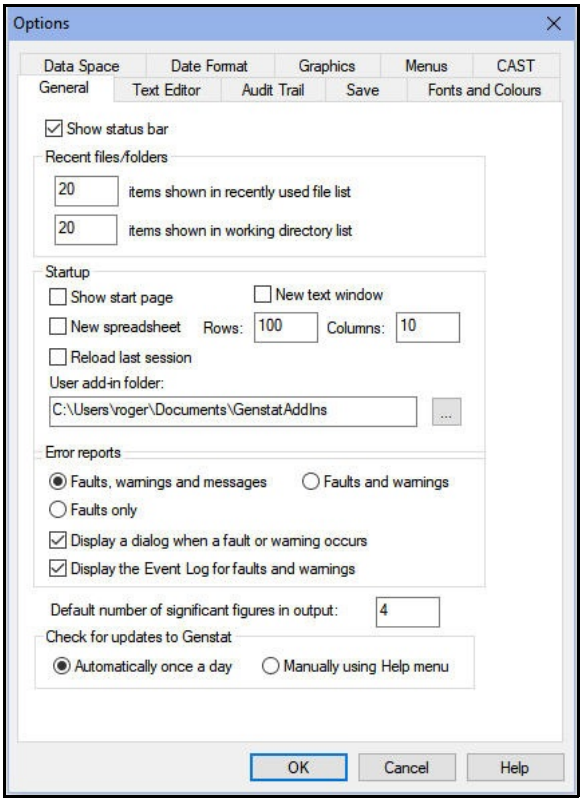


Figure 1.37

1.8 Practical

Open the Options menu and explore the Fonts and Colours tab, which allows you to change the fonts and colours used in the various windows.

2 Data input and calculations

The first step in any data analysis with Genstat (as with any computer software) is to import some data. You can access data from various types of data files on the computer: spreadsheet files (e.g. Microsoft Excel), database formats, simple text files (ASCII files), or files previously prepared by Genstat. You can also use the standard Windows™ cut-and-paste facilities to transfer data directly from other Windows™ applications via the clipboard. We will show how you can import data from Excel by reading from files and by using the clipboard.

Finally, we will introduce the calculations menu, which provides a very powerful calculator that operates with any type of numerical data structure.

2.1 Genstat data structures

In Chapter 1 we introduced two different types of data structures and how these could be read into Genstat using an existing Genstat spreadsheet. When a data set is read into Genstat it is stored within a central data pool, and information on current data structures can be viewed in the Data View pane. The first data structure we introduced was a *variate*, which stores a column of numerical values. The length (or number of values) of a variate is fixed, and two variates of different lengths cannot be used in a common calculation (unless you are calculating summary statistics from them). The second data structure was called a *factor*. A factor is a special data structure within Genstat for specifying an allocation of units into discrete groups. Each group can be represented with a label and/or a numerical value (level). The groups are also assigned ordinal values that are numbered from 1 upwards, and these indicate the order the levels or labels of the factor will be displayed. For example, the table below shows the 3 attributes of a factor that has 4 groups:

Ordinal	**Level**	**Label**
1	0.0	Control
2	0.5	Half Rate
3	1.0	Standard
4	2.0	Double Rate

The levels and labels of a factor can be reordered, but the ordinal values are always numbered 1,2,3… and the order of these cannot be changed. For example, in the table below, the labels and levels have been reordered by sorting the labels alphabetically.

Ordinal	**Level**	**Label**
1	0.0	Control
2	2.0	Double Rate
3	0.5	Half Rate
4	1.0	Standard

When data are imported from other file formats, Genstat will default the order of the levels of any factor either numerically for levels, or alphabetically for labels. You can control the way factor labels and levels are imported using the spreadsheet options and options on dialogs when loading data. The spreadsheet facilities provide a number of menus to allow factor levels and labels to be manipulated. Details are in the *Guide to the*

Genstat Spreadsheet, which can be opened by clicking on the Spreadsheet Sub-option of the Genstat Guides option of the Help menu on the menu bar (see Figure 1.6).

There are many other data structures available within Genstat, each with appropriate attributes. A single numerical value is stored within a *scalar*. A column of textual information is contained in a *text*. A two dimensional array of data is contained in a *matrix*, and the two specialized forms of matrices (*symmetric* or *diagonal*) can also be used. Numerical results of cross tabulations or analyses are stored in *tables* that are indexed by a number of classifying factors.

2.2 Data input

One important feature of Genstat is that it provides flexibility to make tasks easier. A good example of this is data input, where there are three different ways of importing data. The first, and most common way, is to use the File menu (this is a standard Windows™ approach). Opening a spreadsheet or database file through the File menu will load the data into a Genstat spreadsheet for viewing and/or editing before any updating of the data in Genstat's central data pool. The second approach is to open spreadsheet or database files using the Load option of the Data menu. Opening data this way will read the data direct into Genstat's central data pool without displaying the data into a spreadsheet. The final method is to use the Spread menu, which is similar to the File menu where data is opened within a spreadsheet. The Spread menu is useful for creating blank spreadsheets, copying data from the clipboard and viewing data already contained within the central data pool; see the *Guide to the Genstat Spreadsheet* for details.

To illustrate the facilities for data input within Genstat, we will demonstrate how to open data from Microsoft Excel. When reading data from a foreign file, Genstat expects the data to be in a rectangular column format. In a spreadsheet, such as Microsoft Excel, the data need to be arranged in a group of columns forming a rectangle where the columns are of the same length. If the rectangular area contains empty rows or columns then, by default, these will be removed when the data is opened in Genstat. You can specify column names for your data by entering a label for the name in the first row of the column within the rectangular block. A spreadsheet column name must start with a letter (A-Z, a-z or %) and can only contain letters, numbers or the symbols % and _. When data are read into Genstat, a check is made to see if a column name meets these conditions and modifies any names that include invalid characters. For example, if the first character of the column name is a number, then Genstat will create a new name by prefixing the label for the column with a %. When no column names are provided, Genstat will generate default column names using the notation C1, C2 etc... You can specify missing data values by either leaving the cells blank or by entering an asterisk (*).

When the data columns are read into Genstat, any numerical columns will be imported as variates and any column containing labels (excluding the column name) will be imported as a text data structure. Within a Genstat spreadsheet a text column is marked by a green 'T' next to its column name and the contents are right justified by default. A column of numbers or text can also be read into Genstat as a factor. You can specify a column to be a factor by appending an exclamation mark (!) onto the column name (e.g. crop!).

	A	B	C
1	counts	crop!	
2	18	pea	
3	117	pea	
4	21	cereal	
5	7	pea	
6	176	cereal	
7	85	cereal	
8	244	cereal	
9	4	pea	
10	55	cereal	
11	8	pea	

Figure 2.1

Figure 2.1 shows an example of a block of data within the `Genstat Data` worksheet of the Excel file `Bacteria.xls`, which has been arranged for input into Genstat. The data values are a set of counts from an experiment: the numbers of one particular type of bacteria found in small samples of soil growing two different types of crop. The second column contains categorical data and has had the symbol '!' appended to the column to specify the column is to be a factor.

We now look at the first method of importing data into Genstat; using the File menu. In this example we want to open the Excel file containing the data shown in Figure 2.1. To open the file we select the Open line in the File menu on the menu bar. This opens the Open file menu (Figure 2.2), in which we have selected Other Spreadsheet Files from the drop-down list entitled Files of type.

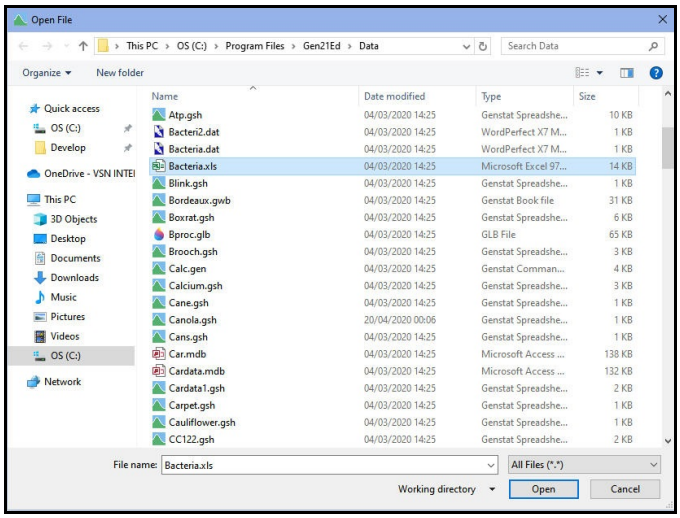


Figure 2.2

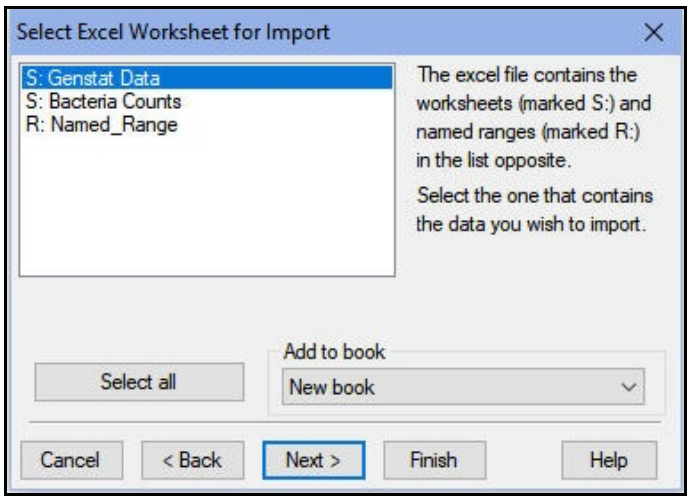


Figure 2.3

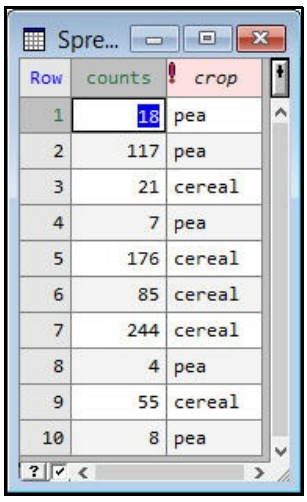


Figure 2.4

Selecting the file `Bacteria.xls` and clicking on Open, or double-clicking on the filename, gives the menu shown in Figure 2.3. This is the initial menu of a wizard for the input of data from an Excel file. It lists all the available worksheets and named ranges within the Excel file, with worksheet names prefixed by 'S:' and named ranges by 'R:'. In this example, we have selected the worksheet Genstat Data. We have no other books or spreadsheets open within Genstat, so the Add to book drop-down list is left as New book. We will explain how to form books of spreadsheets in Chapter 4. Until then, we will keep our spreadsheets separately.

Subsequent menus allow you to select ranges and columns, and set various other options controlling how the data are transferred to Genstat. In this case we want to take all the data on the page, and will leave the other options with their default settings. (The subsequent menus will be shown later though; see Figures 2.7, 2.8 and 2.9.) So we click on Finish to open the two columns of data into a Genstat spreadsheet, as shown in Figure 2.4.

If you click on the Output window, the data in the spreadsheet are automatically transferred to Genstat's central data pool. The Output window displays a brief summary of the data that have been transferred, as shown below.

```
Data imported from Excel file: C:\Program Files\Gen21Ed\Data\Bacteria.xls
 on: 7-Sep-2020 10:41:42
 taken from sheet "Genstat Data", cells A2:B11

Identifier      Minimum         Mean      Maximum       Values      Missing
counts            4.000        73.50        244.0           10            0

Identifier       Values      Missing       Levels
crop                 10            0            2
```

In fact, whenever you change from the spreadsheet window to another window, Genstat will update the central pool with any changes that you have made in the spreadsheet. You can verify that the data have arrived, by looking in the Data View pane (Figure 2.5).

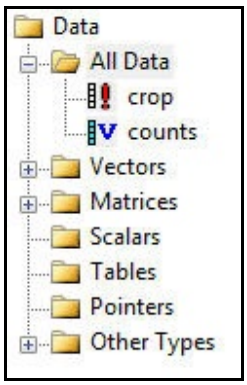


Figure 2.5

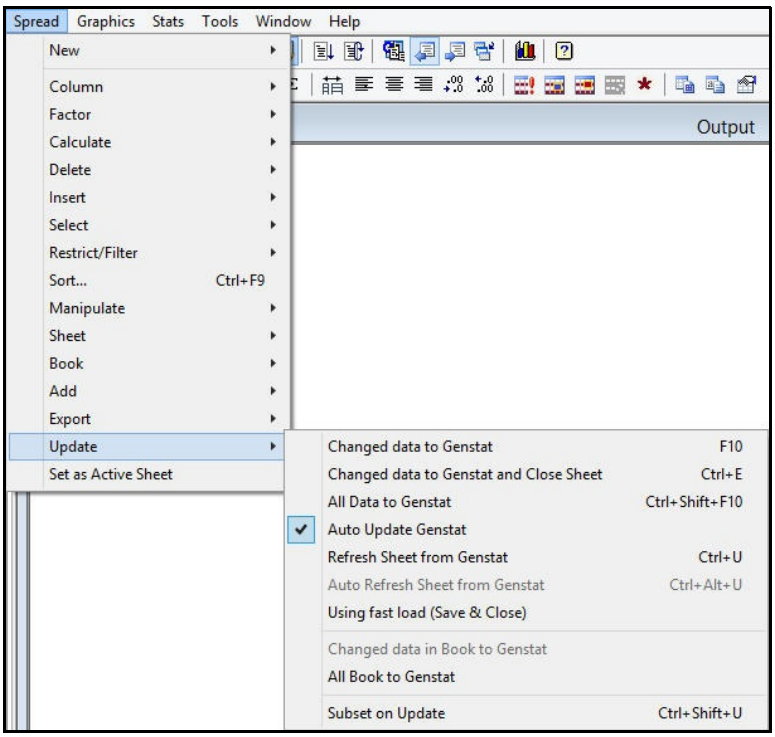


Figure 2.6

You can transfer the data explicitly, using an option from the Update submenu of the Spread menu (Figure 2.6). For example, selecting the Change data to Genstat and Close Sheet item updates Genstat and then closes the spreadsheet. The standard method of updating the pool uses the Genstat `READ` command (see Section 2.15). The Using fast load (Save & Close) item provides a more efficient alternative, using the `SPLOAD` command, for large spreadsheets in Genstat's own `GSH` format (see Section 2.15).

We shall now close the Genstat spreadsheet, and input some data from the other Excel worksheet.

Data are not always stored in a singular rectangular format within a spreadsheet, but may have multiple blocks of data entered on a single worksheet. Figure 2.7 shows an example of this in the worksheet Bacteria Counts from the file `Bacteria.xls`. In this worksheet there is a title in row 1 of column A, and two rectangular sets of data records. In this example we just want to open the second rectangle of data (`counts2` and `crop2`) within a spreadsheet.

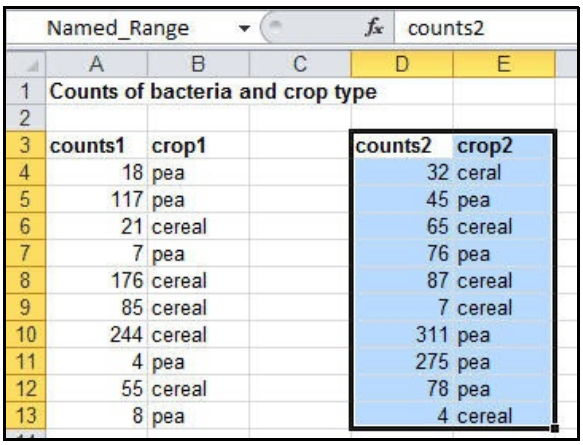


Figure 2.7

This time we shall use the second method of importing data, with the Data menu. This uses the same menus as those with the File menu. However, the data are loaded directly into Genstat's data pool, and no Genstat spreadsheet is formed.

To load the data from the file `Bacteria.xls`, we select the Data File option from the Load item on the Data menu. This produces the Select Input file menu (Figure 2.2) again. So, as before, we select Other Spreadsheet Files from the drop-down list entitled Files of type, select the file `Bacteria.xls`, and click OK. This again leads to the initial menu of the Excel wizard, as shown earlier in Figure 2.3.

There are two ways of reading a rectangular range of data from Excel into Genstat. If we select the worksheet Bacteria Counts in Figure 2.3 and click on the Next button (instead of Finish), the second menu in the wizard, allows the range to be specified explicitly. You check the Specified range radio button (instead of the default All cells), and enter the range D3:E13 into the adjacent field as shown in Figure 2.8.

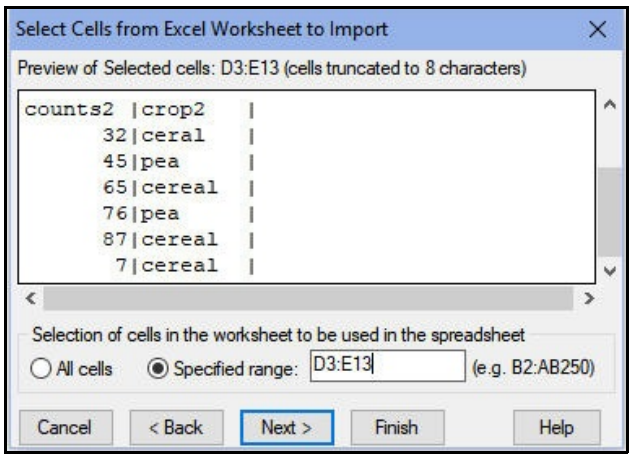


Figure 2.8

Alternatively, you can create a named range for the rectangular block of data within Excel and select this from the worksheet list in Figure 2.3. To create a named range in Excel you first select the desired rectangle either with the mouse or by using the shift and cursor keys. Once the rectangle has been selected, you can name the range by clicking in the Name Box and typing its name. In Figure 2.7 we have selected the range D3 to E13 and entered its name as Named_Range in the Name Box. If you select Named_Range as the worksheet or range in Figure 2.3 and again click Next, you will see that the range D3 to E13 is set up automatically in the second menu of the wizard, just as in Figure 2.8.

The third menu in the wizard (Figure 2.9) allows you to choose which of the columns in the worksheet or range to read. By default they are all read.

The final part of the wizard, shown in Figure 2.10, is a menu with tabs controlling more advanced aspects. This time we have not put an exclamation mark at the end of the column name to specify that the column `crop2` is to be a factor.

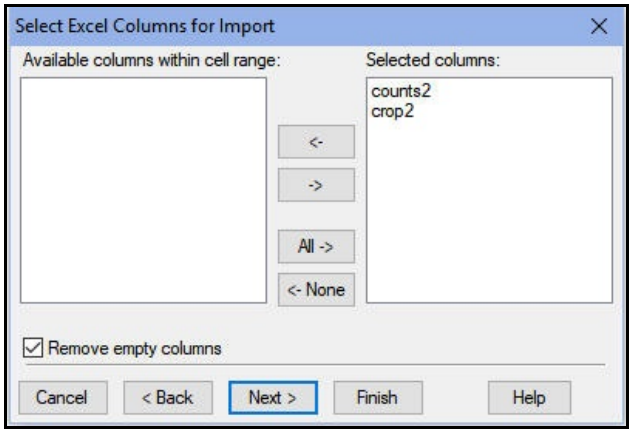


Figure 2.9

So, instead, we select the Factors tab, and check the Suggest columns with only a few unique values to be Factors box. If this option is set, Genstat will check all the columns for repeated values or labels and, if any are detected, you will be prompted with a menu offering you the choice to convert them.

Figure 2.10

On clicking Finish Genstat detects that the column `crop2` has repeated labels and displays the menu shown in Figure 2.11. This menu displays all the columns that have repeating values and the current data type for each column is indicated by a prefix to the name (T specifies a text, F a factor and V a variate). To change the type of `crop2` from a text to a factor we double-click on the name `crop2` in the list (alternatively you can click on the button labelled Factor). This changes the prefix from T to F specifying the column will be a factor. Clicking on OK loads the data range direct into the data pool and produces the summary below.

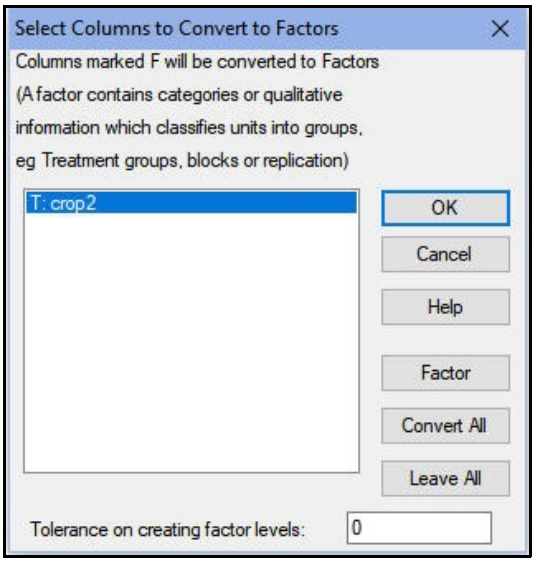


Figure 2.11

.

Data imported from Excel file: C:\Program Files\Gen21Ed\Data\Bacteria.xls
on: 7-Sep-2020 10:44:16 taken from sheet "Bacteria Counts", cells D4:E13

Identifier	Minimum	Mean	Maximum	Values	Missing
counts2	4.000	98.00	311.0	10	0

Identifier	Minimum	Mean	Maximum	Values	Missing
crop2				10	0

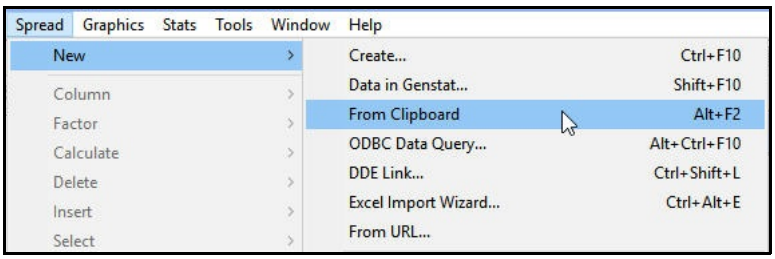


Figure 2.12

The final way to input data is to use the facilities within the Spread menu. In this example we will copy the columns `count1` and `crop1` from the file `Bacteria.xls` (see Figure 2.7) via the clipboard into a Genstat spreadsheet. As with the layout within a spreadsheet, Genstat expects the data on the clipboard to be in a rectangular format with columns of equal lengths. In Excel we select the data, including the column names (data range A3:B13), and then select Copy from the Edit menu. Note that when you are using Excel, if you do any other operation on the spreadsheet before going to Genstat, Excel clears the data from the clipboard. The data is available to Genstat only while the dotted lines are moving around the selected cells in Excel. Now, in Genstat, we create a spreadsheet of the data, by selecting the From Clipboard item from the New option on the Spread menu as shown in Figure 2.12. The New Spreadsheet from Clipboard menu (Figure 2.13) is then produced to control the process. We leave the Suggest columns to be factors box checked (and again ask to open a New Book). Genstat displays the factor conversion menu again. This time it will show `crop1` as the column with repeated values rather than `crop2`, as in Figure 2.11. Leaving `crop1` as a text and clicking OK produces the spreadsheet shown in Figure 2.14.

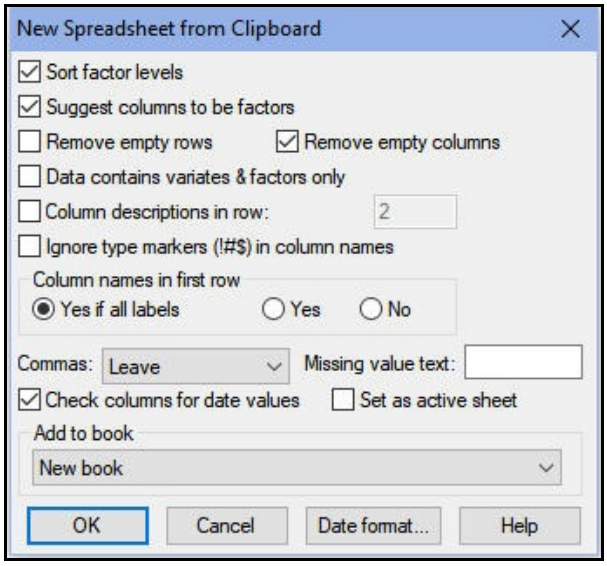


Figure 2.13

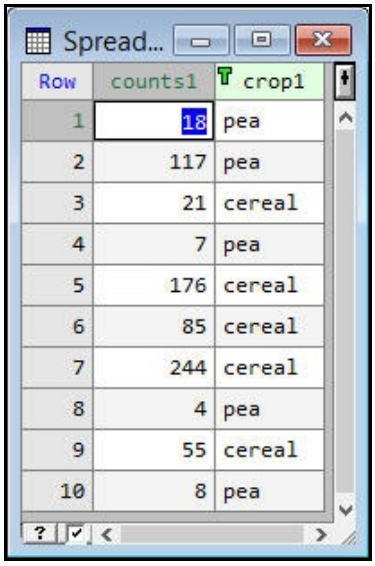


Figure 2.14

2.3 Practical

The file `Traffic.xls` is an Excel data file with one worksheet called counts storing one set of data in the area B3:D43. Use the File menu to read the data into Genstat, converting `day` and `month` to factors. Examine the distribution of the counts using a histogram. (Hint: use the Summary Statistics menu).

2.4 Reading data from ASCII files

In Section 2.2 we described how data can be imported into Genstat in file formats from other applications such as Excel. However, there is another common way to record data in files in a form that can be understood by any application. This is to store the data in a flat (or ASCII) file. This type of file can easily be read and written by Genstat. Data in ASCII format are imported using the Data menu.

The data that we have examined so far in this chapter are just a part of a larger experiment. The computer file `Bacteria.dat` has 40 records, starting with the ten counts we have already looked at, and continuing with a further 30. So, the file looks like this:

```
"Counts of bacteria and crop type, Jan-Feb 2008"
 18 pea
117 pea
 21 cereal
  7 pea
...
  1 cereal
 12 cereal
  4 pea
  * cereal
```

Select Load and then ASCII file from the Data menu (as shown in Figure 2.15). This option can cope with any single set of data laid out in a rectangular array, with columns for measurements and rows or records for observations. The values must be separated from each other by spaces or some other consistently used character such as a comma. Comments can be included in the file by placing a double-quote (") at the beginning of a line. Textual values must be delimited by single quotes (' ') if they contain a space or a comma, or a backslash character (\) which is the continuation character in the Genstat command language.

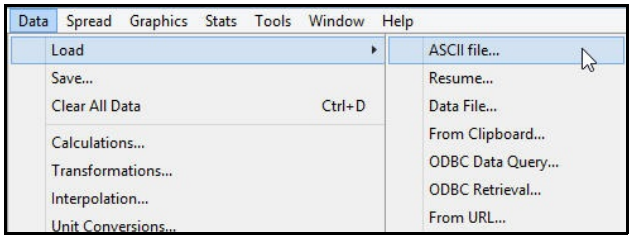


Figure 2.15

The ASCII file option displays the Read Data From ASCII File menu in Figure 2.16. You need to enter the name of the file storing the data, into the ASCII data filename box. If the file is in the working directory (as shown on the status bar), the filename is sufficient. Otherwise, you can change the current directory, or give the full path of the file. It is probably easiest, however, to click on Browse to find the directory and file that you want. Figure 2.17 shows the Open

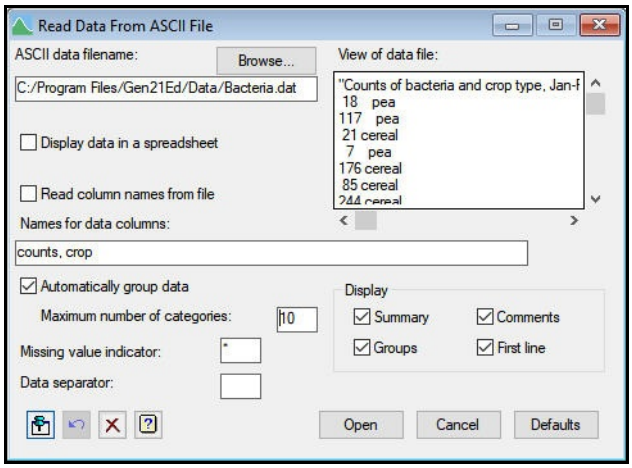


Figure 2.16

File menu, with the required file highlighted.

Double-clicking on the file Bacteria.dat will enter the full filename into the box in the Read Data from ASCII file menu, and display the first five lines of the file in the box at the top right of this menu.

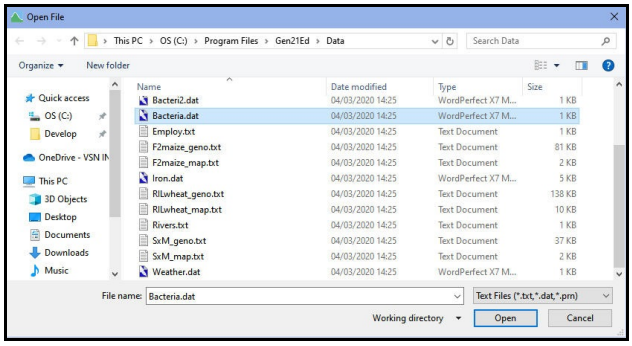


Figure 2.17

The names for the data columns are typed in the Names for Data Columns box, separated either by spaces or commas; there must be one name for each column of data in the file.

You can choose to group the data automatically to form a *factor*, by checking the box Automatically group data, as in Figure 2.16. The grouping option works by checking the number of distinct values in each column of data, and querying whether to turn any column that has 10 or less into a factor. You can change the value for the grouping criterion by entering a value in the Maximum number of categories box. Now click on the Open button to read the data.

With this set of data, the counts will not be grouped because there are many more than 10 different counts. However, there are only two different crop names, pea and cereal, so the menu shown in Figure 2.18 will appear asking whether to turn crop into a factor or keep it as a text.

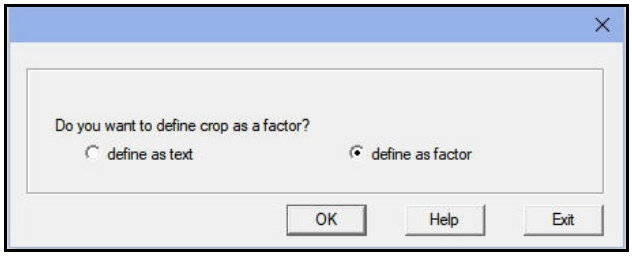


Figure 2.18

Genstat reads the data file and sets up data structures to store the values that it finds. A report is sent to the Output window, controlled by the check-boxes at the bottom of the Read Data menu, and this is shown below.

Initial comments

```
"Counts of bacteria and crop type, Jan-Feb 2008"
```

First line of data with no missing values

```
 18     pea
```

Summary

The file C:/Program Files/Gen21Ed/Data/Bacteria.dat is assumed to contain 2 structure(s), with one value for each structure on each record.

Occurrence of distinct values of crop

	Count
category	
cereal	20
pea	20

The file contains 40 values for each of the following structures:

Identifier	Type	Missing
counts	variate	3
crop	factor	0

There are several other options in the Read Data from ASCII file menu, that make it easier to deal with data that are not in Genstat's standard form. To illustrate these, we have formed another file called `Bacteri2.dat` which contains the same data, but with comma separators and the minus sign (`-`) rather than the asterisk (`*`) as the missing-value indicator. The identifiers for the two data structures are now included in the first line of the file. Figure 2.19 shows the Read Data from ASCII File menu, completed, ready to read the data from this file instead. Notice that it can save time to store identifiers for data structures in the file with the data; whenever you access the file from Genstat, you will not have to remember what order the columns are in, nor type the names that you wish to use.

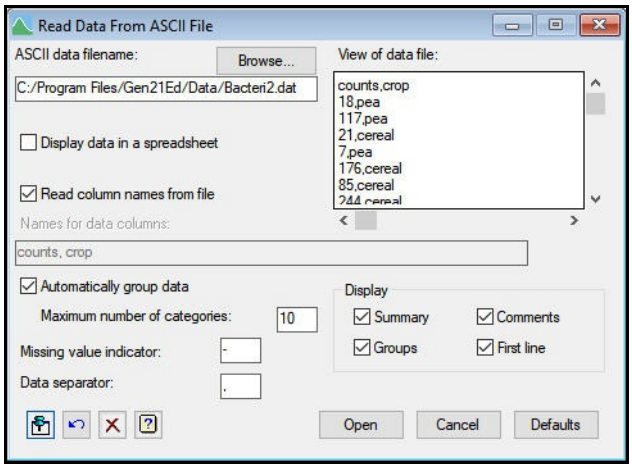


Figure 2.19

2.5 Practical

The file `Rivers.txt` is an ASCII file containing information on the 15 longest rivers.

```
River       Length      Continent
Nile        6695        Africa
...
Amur        4415        Asia
Lena        4269        Asia
Mackenzie   4240        N.America
Niger       4183        Africa
Mekong      4180        Asia
Yenisey     4090        Asia
Murray      3717        Oceania
Volga       3688        Europe
```

Use the Data menu to read all the data, storing the third column as a factor. The name Huang He contains a space, so it is enclosed in single quotes in the data file to ensure that it is treated as a single data value rather than two values. Examine the attributes of the data structures that have been read, using the Data View pane (see Figure 1.12).

2.6 Displaying data

There are two ways that you can view data structures from the Genstat central data pool. You can display the contents of data structure either within a spreadsheet or within the Output Window.

To display data in a spreadsheet you can use the Load Spreadsheet menu. The advantage of displaying the data within a spreadsheet is that you are then able to edit the values and send these changes back to Genstat. To open the Load Spreadsheet menu we select the Data in Genstat item from the New option on the Spread menu, which opens the menu as shown in Figure 2.20. In this menu you are offered the choice of several different types of spreadsheet, allowing you to display different kinds of data structures such a vector of values, matrix or table.

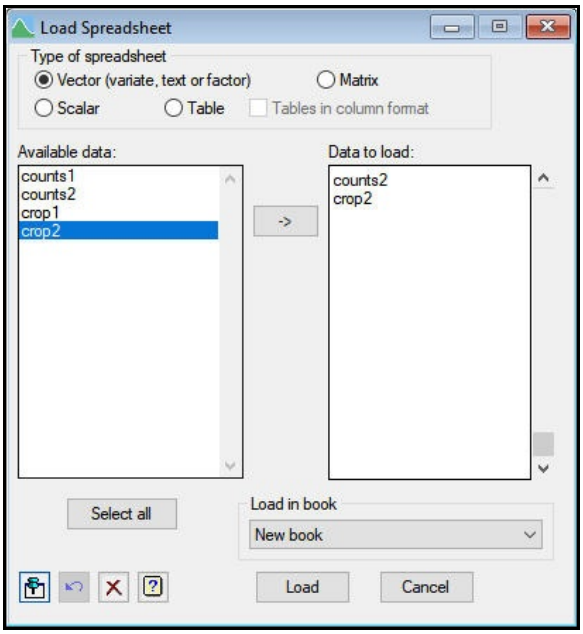


Figure 2.20

In this example we have selected the default *Vector Spreadsheet*, which allows a spreadsheet to be formed using one or more columns (or vectors) of information. A Vector Spreadsheet can contain variates, texts and factors simultaneously if they are all the same length. To select the columns to display in the spreadsheet we have double-clicked the names `counts2` and `crop2` in the Available data list to copy them into the Data to load list.

Clicking on Load produces the spreadsheet shown in Figure 2.21.

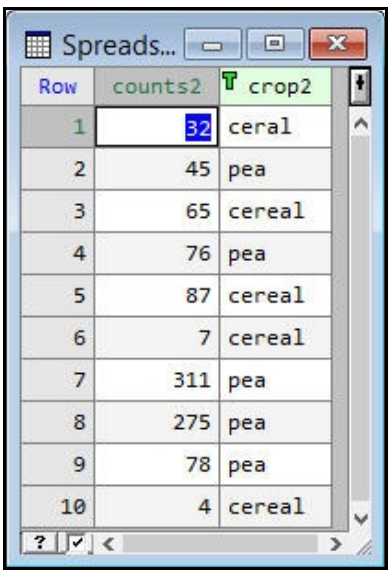


Figure 2.21

To display the contents of data structures within the Output Window you can use the Display Data in Output menu. To open this, click on the Display Data in Output option of the Data menu on the menu bar. The menu is shown in Figure 2.22.

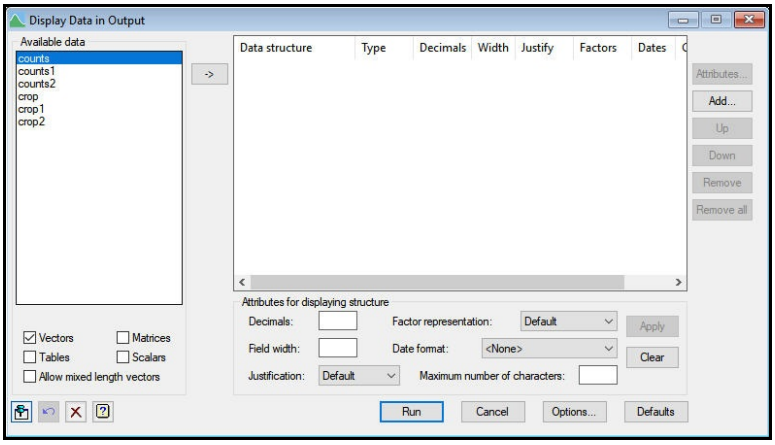


Figure 2.22

For this example we shall illustrate how to display the data structures `counts` and `crop` from Section 2.4. The idea is that you define the formatting for each data structure, using the fields at the bottom of the menu, and then enter it into the right-hand window by clicking on the arrow.

Once we have entered `counts` (setting a field width of 8), the contents of the Available date box change to offer only the vectors with the same length as `counts`; see Figure 2.23. (We can still display vectors with different lengths, but would need to check the box to Allow mixed length vectors first.) We then enter `crop`, keeping a field width of 8.

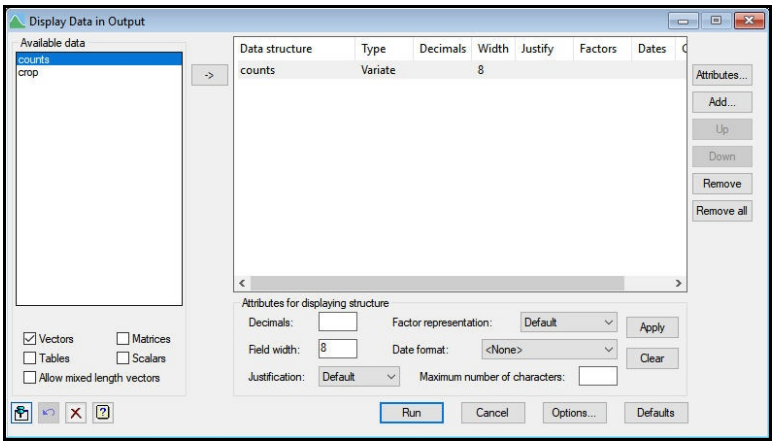


Figure 2.23

You can modify the display format for a particular dat structure, by highlighting its line in the right-hand window, and then clicking on the Attributes button to open the Display Attributes menu, as shown in Figure 2.24. Alternatively, you can modify the settings in the boxes at the bottom of the menu, and click on Apply to use those instead of the existing settings.

Once we are happy with all the settings, we can click on Run to display the data in the Output Window as shown below.

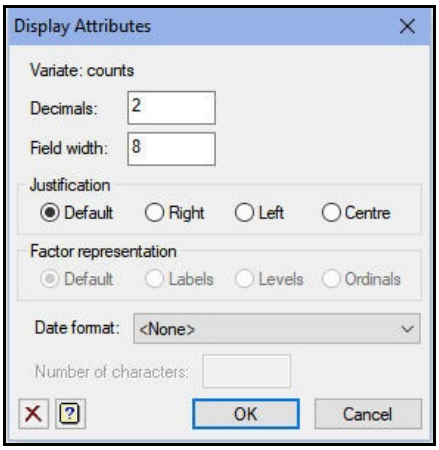


Figure 2.24

counts	crop
18.00	pea
117.00	pea
21.00	cereal
7.00	pea
176.00	cereal
85.00	cereal
244.00	cereal
4.00	pea
55.00	cereal
8.00	pea
73.00	cereal
*	pea
3.00	pea
4.00	pea
40.00	cereal
198.00	cereal
123.00	pea
17.00	pea
74.00	cereal
3.00	pea
2.00	pea
5.00	pea
0.00	cereal
4.00	pea
2.00	pea
10.00	cereal
2.00	cereal
4.00	cereal
3.00	pea
1.00	pea
*	cereal
2.00	cereal
4.00	pea
1.00	pea
4.00	cereal
15.00	cereal
1.00	cereal
12.00	cereal
4.00	pea
*	cereal

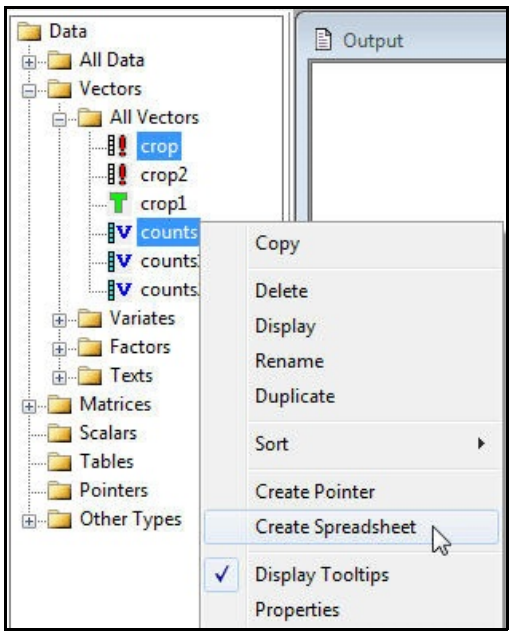


Figure 2.25

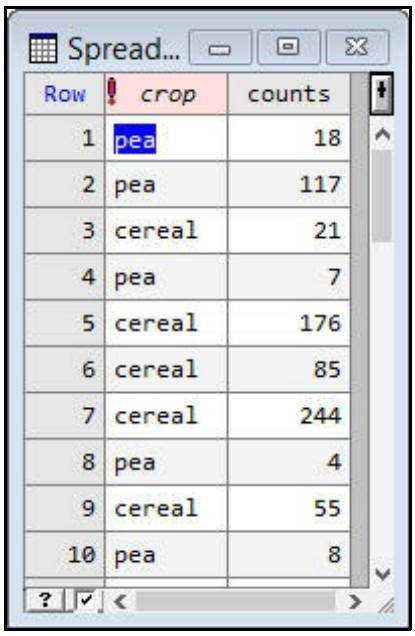


Figure 2.26

You can also use the Data View pane to display data structures within a spreadsheet. For example, to open a spreadsheet containing the data structures `counts` and `crop`, select the two names in the data list as in the previous example using the Control key and left mouse clicks. Clicking on either of the selected names using the right mouse button will produce the menu shown in Figure 2.25. Selecting the Create Spreadsheet menu option will produce a spreadsheet (Figure 2.26) containing the two selected data structures.

2.7 Practical

Use the Display Data in Output Window menu (Figure 2.23) to display the names and lengths of the rivers from Practical 2.5.

2.8 Converting data structures

After reading data into Genstat you may want to change the data type. For example, you may want to convert a column to a factor as it contains grouped data or a factor to a text for use in a particular menu. To illustrate how to convert structures we will change the column `crop` from Figure 2.26 into a text using the menus and change it back to a factor using the right-mouse short-cut menu. So, to convert the column `crop` to a text structure, we select Convert from the Column item on the Spread menu.

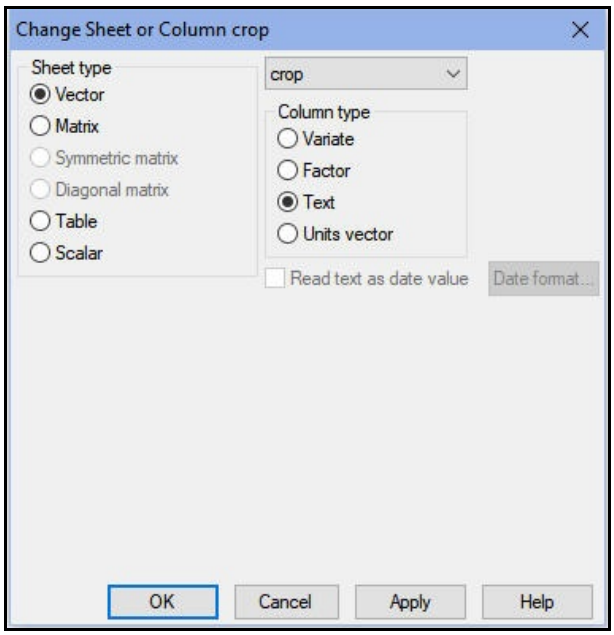


Figure 2.27

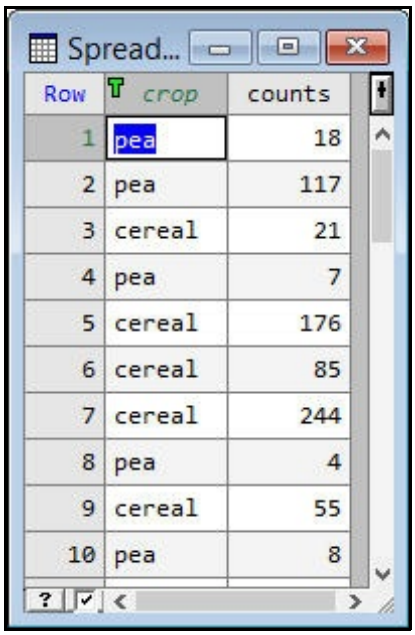


Figure 2.28

This opens the menu shown in Figure 2.27. The different types of data structure to which you can convert are listed in the Column type box. We have selected `crop` from the list of columns and we have selected Text from the Column type box. Clicking OK on this menu changes the spreadsheet as shown in Figure 2.28. You can now see that the column heading has a green 'T' indicating that the column is a text structure.

To change the column back to a factor, move the mouse over the column `crop` and click the right-mouse button. This should pop up the menu shown in Figure 2.29. On this menu we have selected the item Convert to Factor. This returns the spreadsheet to its original state as shown in Figure 2.26.

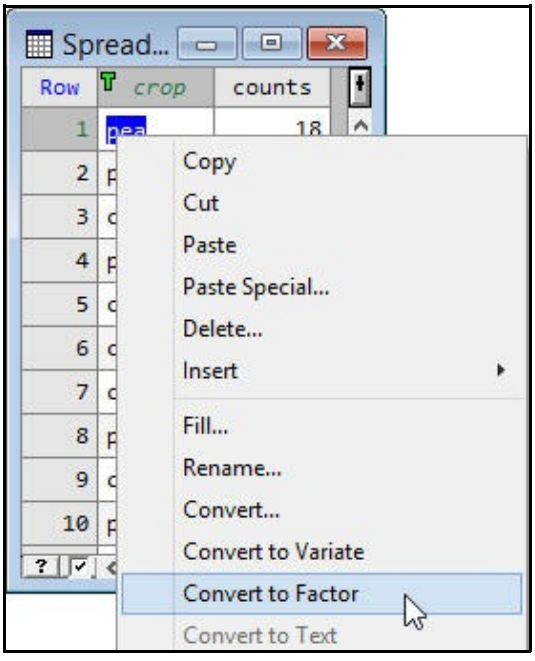


Figure 2.29

2.9 Practical

Load the names, lengths and continents of the rivers from Practical 2.5 into a spreadsheet. Convert continent into a text, and then back to a factor.

2.10 Saving data to files

The easiest way to save data in Genstat to an external file is to display the data within a Genstat spreadsheet, and save this as a Genstat Spreadsheet file (`*.gsh`). To save such a file, select Save As from the File menu. This prompts you with a menu, as shown in Figure 2.30. In the menu select the Genstat Spreadsheet (*.gsh) option from the Save as type list. You can specify the name of the file in which to save the spreadsheet, and select the directory in which to store the file.

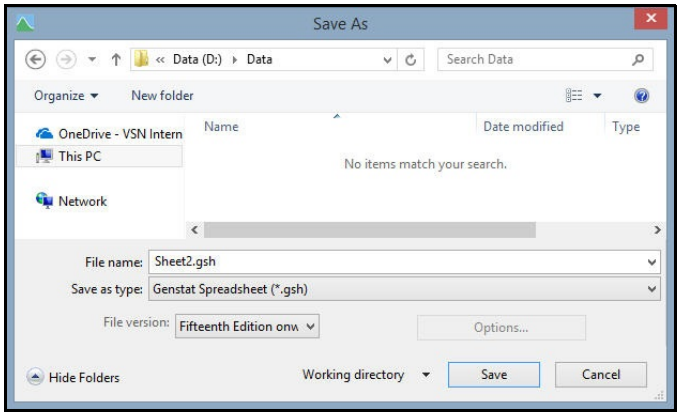


Figure 2.30

Genstat has the ability to save data to files in several other common file formats. For example, you can save a spreadsheet in an Excel file, Lotus workfile or a database file. When Genstat saves the data to these file formats, the data are stored in exactly the same format as they are displayed in the spreadsheet. Storing the data in this way has the advantage that it is easily read back into Genstat. To illustrate this, we look at how the Bacteria counts spreadsheet is saved to an Excel file.

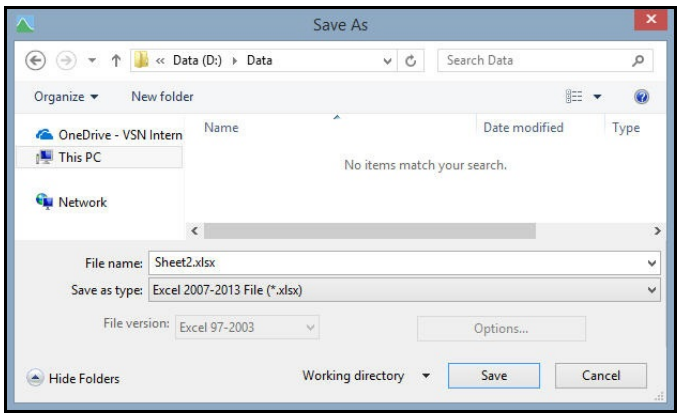


Figure 2.31

To save the spreadsheet, select Save As from the File menu. In the Save As menu select the Excel 2002-2010 File (*.xlsx) option from the Save as type drop-down list from the File version drop-down list (see Figure 2.31). Type in the file name (here the menu shows the default `Sheet2.xlsx`: this is the second unnamed spreadsheet that we have formed in this session), and click Save.

When Genstat saves a new spreadsheet (i.e. one that has not been saved previously) to an Excel file, the data are written to the first worksheet. Figure 2.32 shows the layout used to represent the bacteria data in the Excel File `Sheet2.xls`. The data are stored in exactly the same column format as within the Genstat Spreadsheet (see Figure 2.26), but with the identifier names in row 1, and the data values in rows 2 onwards. The identifier name for the column crop has an exclamation mark, "!", appended to the end of the name. This has been added to the name to indicate that the column of data is of type factor.

	A	B
1	crop!	counts
2	pea	18
3	pea	117
4	cereal	21
5	pea	7
6	cereal	176
7	cereal	85
8	cereal	244
9	pea	4
10	cereal	55
11	pea	8
12	cereal	73

Figure 2.32

Another way of saving data is to save all the contents from the central data pool to a Genstat *Save* File (`*.gsv`). To do this, select the Save option from the Data menu on the menu bar. This brings up the Save data in menu, shown in Figure 2.33, allowing you to select the file in which to store all the current data. A Genstat Save File can be opened at a later date using either the File Open option or the Resume item from the Load option on the Data menu.

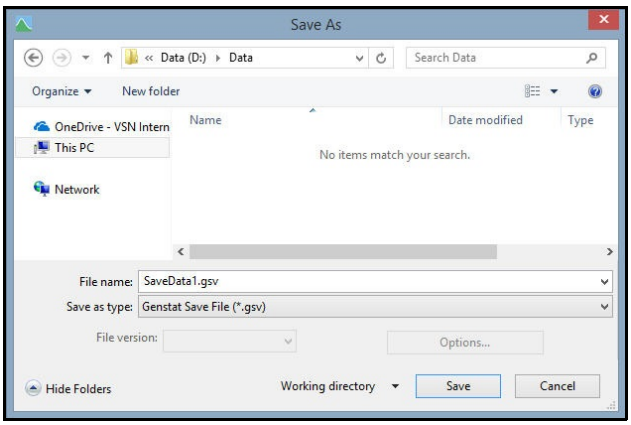


Figure 2.33

There is a special file format in Genstat called a *Session* file (*.gsn), which allows you to save the entire contents of your current Genstat session including the data, spreadsheets, text windows and any open menus. Opening a session file will automatically re-open all of these items. To save a session, select the Save Session option from the File menu. This opens the save dialog as shown in Figure 2.34.

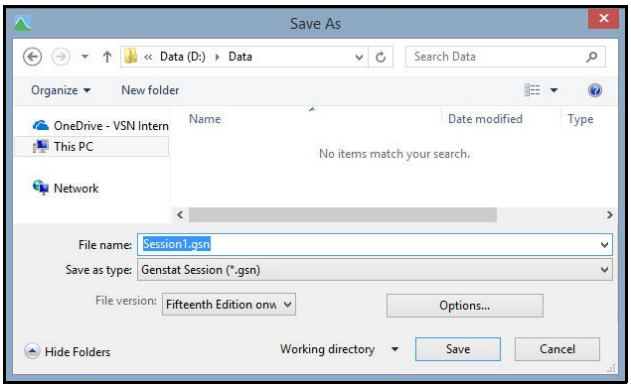


Figure 2.34

You can control what details to save by clicking on the Options button. This opens the menu shown in Figure 2.35 where you can select from a range of different features.

Later parts of this chapter use some new data sets. So we now reinitialize Genstat, clearing the data store and closing any open menus and spreadsheets, by clicking on Run on the menu bar and choosing the Restart Session option.

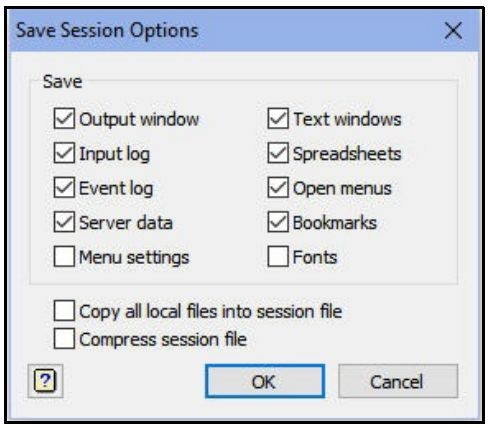


Figure 2.35

2.11 Practical

Save the spreadsheet from Practical 2.9 as an Excel file.

2.12 Calculations

The Calculate menu (Figure 2.36) is obtained by selecting Calculations from the Data menu. The Calculate menu can be used for calculations such as transforming data, and data summaries. It can also work as an ordinary calculator. For example, you can simply multiply two numbers together: type `4`, click the button for the operator `*`, and then type `6.25`. As you do this, you can see that the expression defining the calculation is recorded at the top of the window. You can display the result in the output window, by selecting the Display in output option and clicking Run to produce the following output.

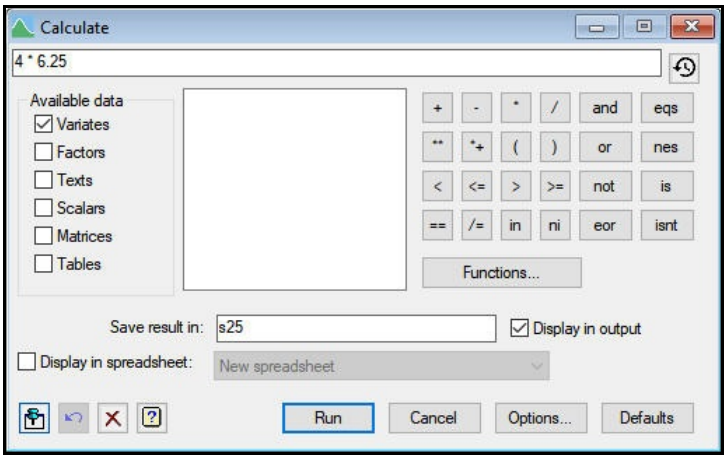


Figure 2.36

```
4*6.25
 25.00
```

Instead of (or as well as) displaying the result, you can ask to save the result in a structure by giving an identifier name, say `s25`, in the Save result in box; `s25` will then be defined as a scalar data structure (since the calculation has generated a single number as its result), storing the value 25.

As you can see, all the usual arithmetic operators are available:

`+` addition; `-` subtraction; `*` multiplication; `/` division;

** exponentiation (for example, X**2 stands for X^2)

There is also the operator *+ for matrix multiplication, left and right brackets, and a full set of logical and relational tests, these are described later.

Most practical calculations are done on whole series of numbers stored in variates. To show what can be done, we shall work with some administrative data from a small company, recording rates of pay and hours of work over a four-week period. These are available in a Genstat spreadsheet file, Pay.gsh, which can be opened by selecting Open from the File menu from the menu bar as described in Section 2.2. If you look at the Calculate menu again (Figure 2.37), the Available Data box will now list the variates hours1, hours2, hours3, hours4 and rate.

Our first calculation, in Figure 2.37, works out the wages for the first week. First clear the menu by clicking on the Clear button (the one with the red cross, third from the left in the line in the bottom left-hand corner of the menu). Now enter the calculation hours1*rate, by double-clicking on the identifier hours1 in the Available data box, single-clicking on the operator button *, and then double-clicking on the identifier rate. Type pay into the Save result in box, and click on the Run button to do the calculation.

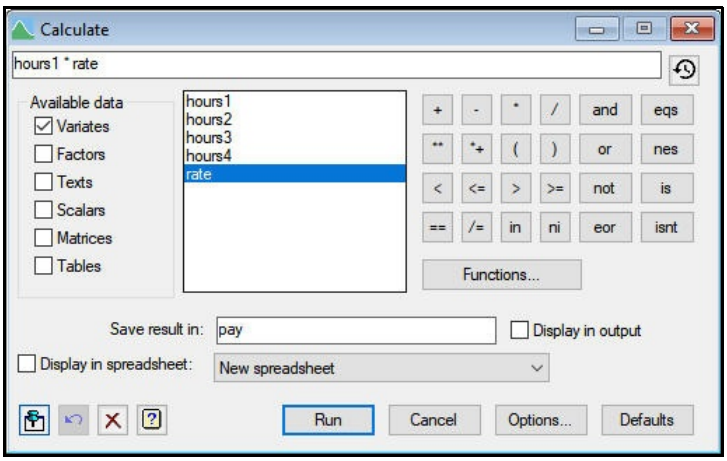


Figure 2.37

The calculation takes place for every unit of the variates and pay is defined as a variate. The value for Foster is the appropriate value for hours1 (41) multiplied by the corresponding value of rate (10.00) and so on. This can be verified by returning to the spreadsheet and adding the calculated column (pay) back into the spreadsheet. Selecting the Data in Genstat item from the Add option of the Spread menu on the menu bar opens the menu shown in Figure 2.38. This menu operates in a similar way to the Load Spreadsheet menu described in Section 2.6 where, instead of creating a new spreadsheet, the data is added as new columns within the current spreadsheet. We have double-clicked the name pay in the Available data list to copy it across to the Data to load box.

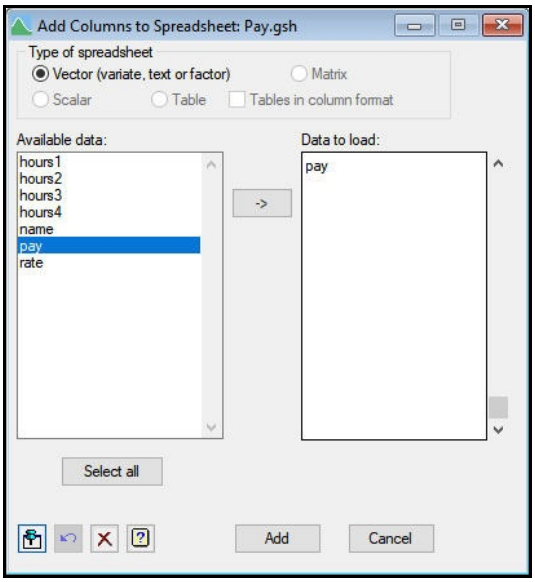


Figure 2.38

Clicking Add produces the spreadsheet shown in Figure 2.39 with the new column pay added to the spreadsheet.

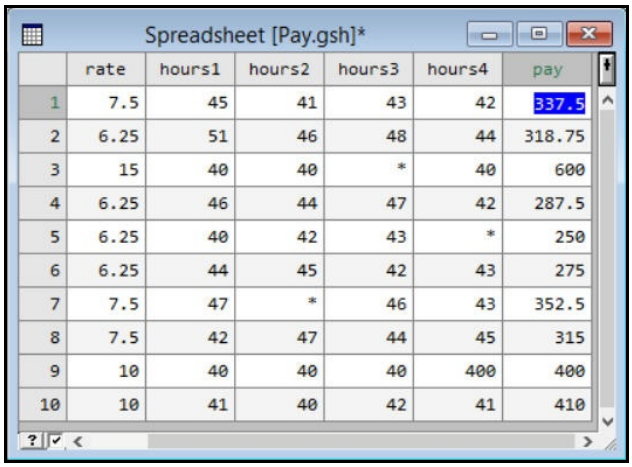


Figure 2.39

You can include a scalar in a calculation with a variate. This will apply the calculation with the scalar value to every unit of the variate. To illustrate this we will add the scalar s25 to every unit of the variate pay. First we click the Clear button again to re-initialize the menu, and check the Scalars box in the Available data box so that the scalar s25 will be displayed in addition to all the available variates. We can then enter the calculation pay + s25, and specify pay1 as the name of the structure (again a variate) to store the results. This time we place the results directly into the spreadsheet. To do this we select the Display in spreadsheet checkbox, which enables a list of open spreadsheets from which we select [Pay.gsh]Sheet1. Figure 2.40 shows the Calculate menu, and Figure 2.41 illustrates the effects of adding pay1 to the spreadsheet where the calculated column is indicated by a yellow block in the column title.

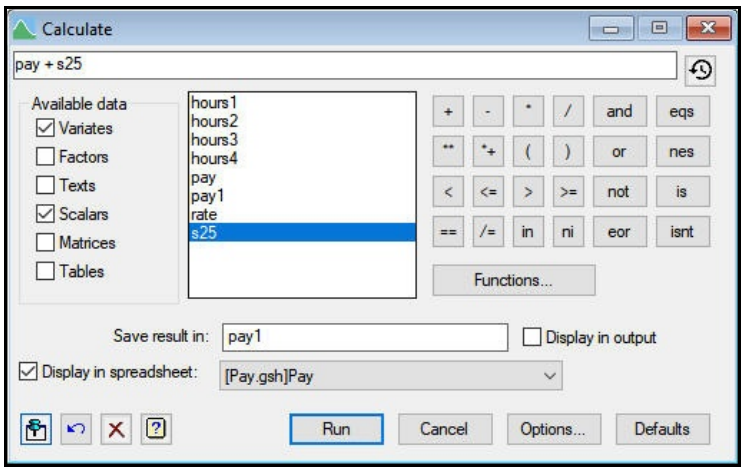


Figure 2.40

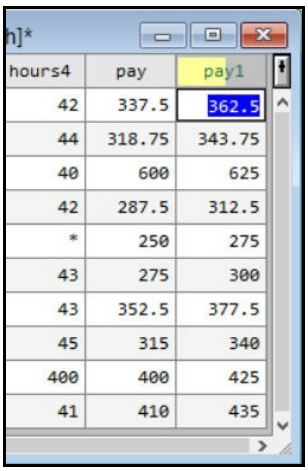


Figure 2.41

Genstat provides a wide range of functions that can be included in the expression. Click on Clear in the Calculate menu to clear the expression and save box. Now, clicking on the Functions button opens the Calculate Functions menu from which you can choose the function and set its arguments, as shown in Figure 2.42.

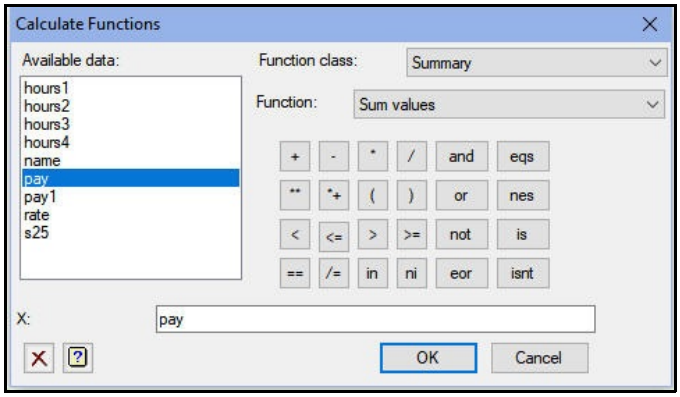


Figure 2.42

The menu contains a drop-down list box of different classes of functions. For each function class there is a range of functions available in the Function drop-down list. In Chapter 3 we will describe how to transform data using the logarithmic function. This type of calculation will produce a result that is the same type of structure as the argument of the function. Classes which produce a result in this form include transformations, inverse transformations, strings, log-likelihood, and the various types of probability. Some function classes however do produce a result that is a different type of structure than that of the argument. For example, the functions in the Summary class produce a scalar summary of all the values in a structure.

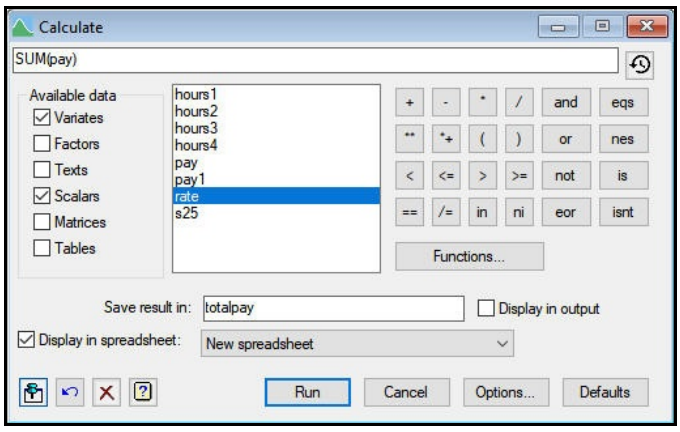


Figure 2.43

To illustrate this we will calculate a new scalar, `totalpay`, as the sum of the pay for all employees in week 1. We select Summary as the class of function, and then Sum values as the function. The Sum values function has a single argument (the numbers to be summed), so we enter this into the X box using the Available data (Figure 2.42). Clicking OK, transfers the function and its argument to the expression box in the main Calculate menu at the current cursor position (Figure 2.43).

Clicking Run in the Calculate menu creates a new scalar spreadsheet with one column for the value for `totalpay` (Figure 2.44).

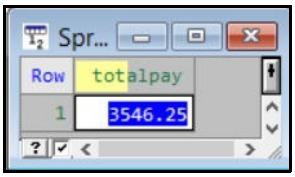


Figure 2.44

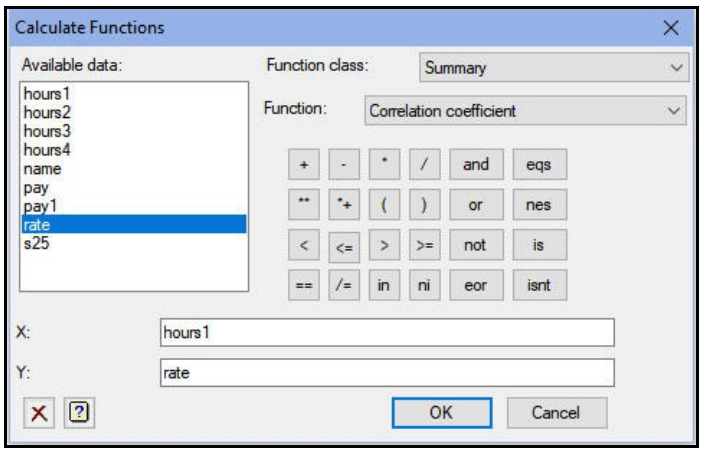


Figure 2.45

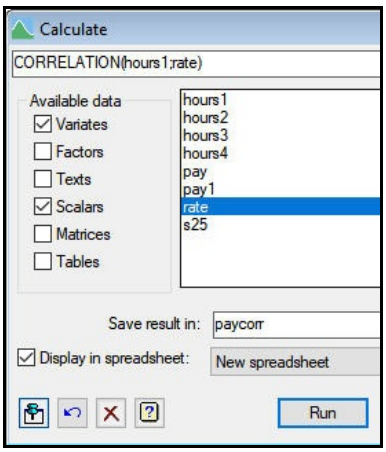


Figure 2.46

Most of the summary functions operate on a single data structure. One exception is `CORRELATION`, which calculates the correlation between two structures. In Figures 2.45 and 2.46, we illustrate how to calculate the correlation between the hours worked in week 1 and the pay rates placing the results in the scalar `paycorr`. The resulting scalar spreadsheet containing the value for `paycorr` is shown in Figure 2.47. Notice that, when a function has more than one argument, each is separated from the next by a semi-colon. (Correlations can also be calculated from the Summary Statistics option of the Stats menu.)

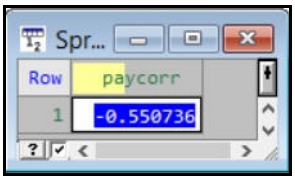


Figure 2.47

Sometimes data will include missing values, for example, the column `hours2` has a missing value in row 7 where there was no record of the hours worked for Edwards in week 2. Genstat has a general rule for calculations involving missing values in that, if any of the structures involved in a calculation has a missing value in a particular unit, the result of the calculation will be missing for that unit. To illustrate this we will calculate the variate `monthpay` as total pay for each employee over the four weeks (Figure 2.48).

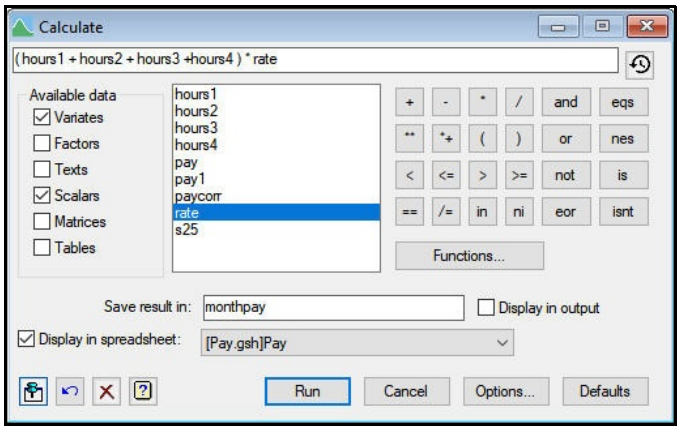


Figure 2.48

The resulting spreadsheet is shown in Figure 2.49 where the column monthpay contains a missing result when the hours are unknown in any of the four weeks. If you wish to replace a missing value, you can use the Replace missing values function (MVREPLACE) in the Transformations menu. This has two arguments: the first specifies the identifier of the data structure with the missing values, and the second supplies the values that are to replace them. In our example we might assume that a value would be missing if an employee had not been present during the week concerned, so we should replace it by zero. To do this we use the calculation shown in Figures 2.50 and 2.51 for each of the hours columns which contain a missing value.

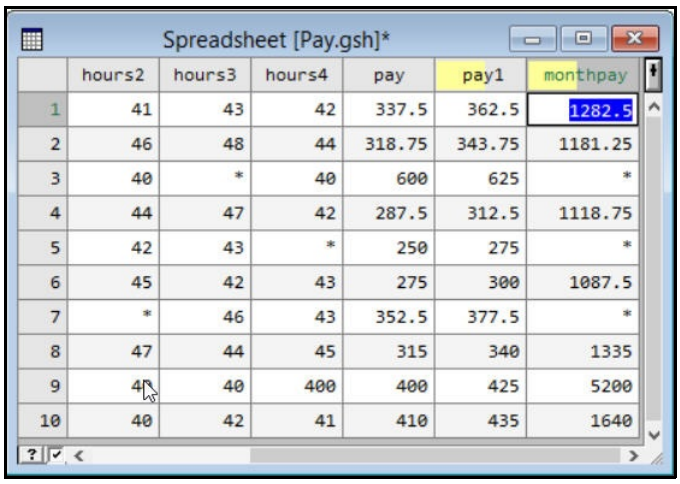


Figure 2.49

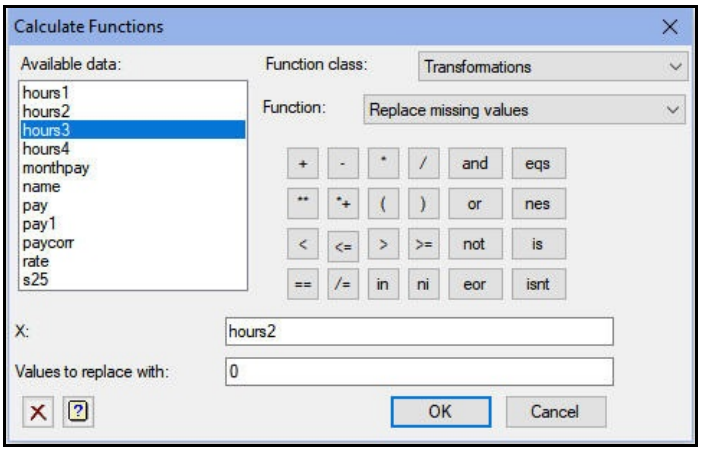


Figure 2.50

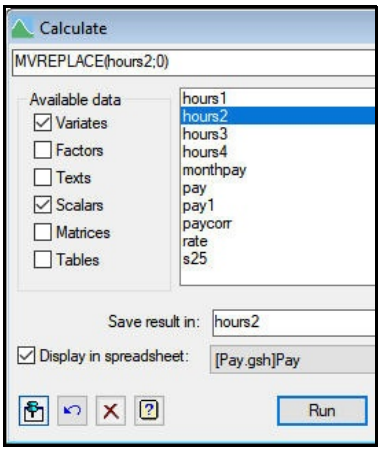


Figure 2.51

When values are changed in data structures that have been used in a calculation, if the calculated structures are present in a spreadsheet (e.g. pay1 and monthpay), they need to be manually recalculated. We can now recalculate monthpay by selecting Recalculate from the Calculate option of the Spread menu, see Figure 2.52. In this menu we select the column monthpay and click OK to recalculate that column in the spreadsheet.

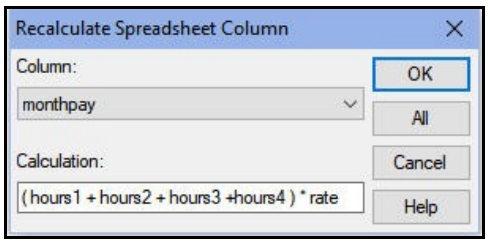


Figure 2.52

If you examine the spreadsheet in Figure 2.49 more closely, you will see that there are other problems in the data: the value in row 9 for `hours4` is 400. Calculations can also involve relational and logical tests. These produce the value 1 if the result is *true*, and 0 if it is *false*. So, in Figure 2.53, we can use the greater-than operator (>) to set up a variate called `odd4` containing 0s and 1s according to whether staff are recorded as working less than or greater than 100 hours in the fourth week.

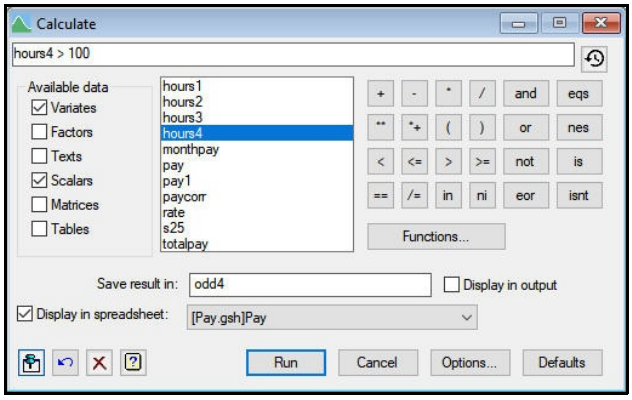


Figure 2.53

We can then use `odd4` in the Insert missing values function (`MVINSERT`) to place a missing value into unit 9 of the `monthpay` variate, since we believe this record must be wrong. This function also has two arguments: the first is the identifier of the structure with values that need changing, and the second is a variate of 0s and 1s indicating which values are to become missing values.

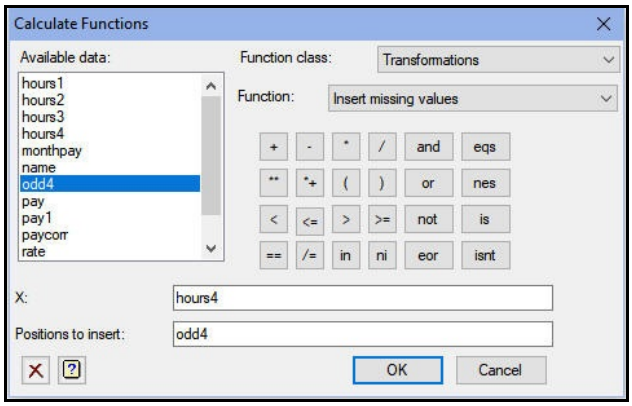


Figure 2.54

So Figures 2.54 and 2.55 take the values of `monthpay`, insert a missing value whenever the corresponding value of `odd4` is non-zero, and store the results back in `monthpay`. The resulting spreadsheet is shown in Figure 2.56.

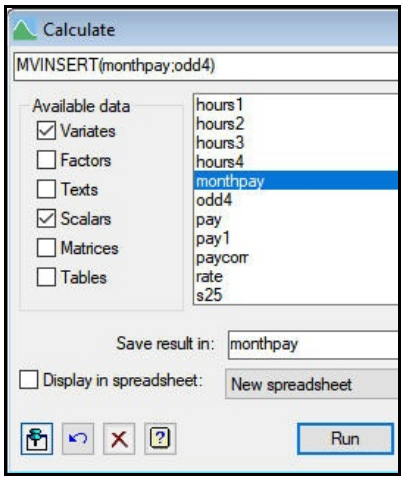
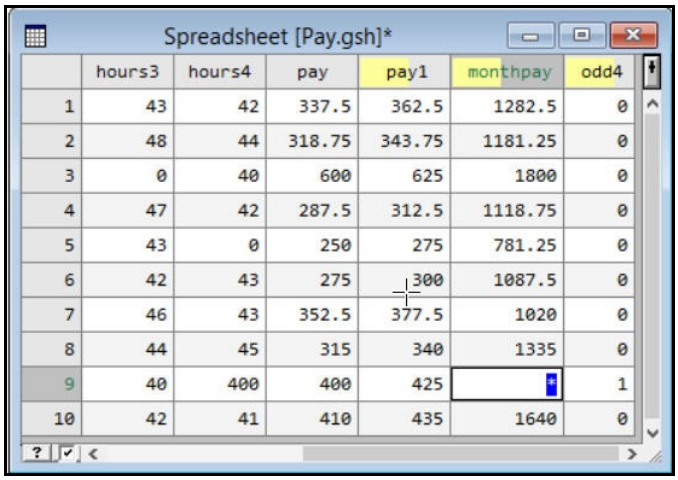


Figure 2.55



	hours3	hours4	pay	pay1	monthpay	odd4
1	43	42	337.5	362.5	1282.5	0
2	48	44	318.75	343.75	1181.25	0
3	0	40	600	625	1800	0
4	47	42	287.5	312.5	1118.75	0
5	43	0	250	275	781.25	0
6	42	43	275	300	1087.5	0
7	46	43	352.5	377.5	1020	0
8	44	45	315	340	1335	0
9	40	400	400	425		1
10	42	41	410	435	1640	0

Figure 2.56

The calculation does not have to be done in two stages: the second argument of the Insert missing values function could be replaced by the expression `hours4 > 100`. However, the intermediate variate `odd4` makes it easier to see what is going on.

The available operators for relational tests are as follows:

`<`	less than
`<=`	less than or equal to
`>`	greater than
`>=`	greater than or equal to
`==`	equality
`/=`	not equal to
`.IN.`	inclusion: `X.IN.Vals` gives result true for each value of `X` that is equal to any one of the values of `Vals`
`.NI.`	non-inclusion: the opposite of `.IN.`

There are also logical operators that can be useful to combine the results of expressions involving relational operators.

`.AND.`	and: `a.AND.b` true if both `a` and `b` are true
`.OR.`	or: `a.OR.b` is true if either `a` or `b` is true
`.NOT.`	not: `.NOT.a` is true for `a` untrue
`.EOR.`	either or: `a.EOR.b` is true if either `a` or `b`, but not both, is true

The precedence of the operators (that is, the order in which they are evaluated if there are several different ones contained in an expression) is much the same as you would expect from ordinary arithmetic:

(1) `.NOT.` Monadic minus (that is, minus as for example in `-2`)
(2) `.IS.` `.ISNT.` `.IN.` `.NI.` `*+`
(3) `**`
(4) `*` `/`
(5) `+` Dyadic minus (that is, minus as for example in `x-y`)
(6) `<` `>` `==` `<=` `>=` `/=` `<>`
`.LT.` `.GT.` `.EQ.` `.LE.` `.GE.` `.NE.` `.NES.`
(7) `.AND.` `.OR.` `.EOR.`
(8) `=` (see the examples of `CALCULATE` below)

Within each class, operations are done from left to right within an expression. So, for example

```
5 - 1 - 2
```

gives the value 2. In case of any doubt, it is safest to use brackets – the expression inside a pair of brackets is always evaluated first. So, for example

```
5 * 2 + 3
```

gives the value 13, but

```
A = 5 * (2 + 3)
```

gives the value 25.

Notice that *monadic minus* is one of the two operators with highest priority. This means that, unless you use brackets, the minus is evaluated before any other arithmetic operator. So, for example, if you want to calculate $-x^2$ you need to put `-(x**2)`.

Genstat *text* structures can be used in expressions, but only with the set inclusion operators `.IN.` and `.NI.` (see above), or the string operators `.EQS.` (equality) and `.NES.` (inequality). For example, the expression

```
text1 .EQS. text2
```

compares the string in each unit (or line) of `text1` with that in the corresponding unit of `text2`, giving the result *true* if they are identical, while

```
text1 .NES. text2
```

gives the result *true* if they differ.

When a factor occurs in a calculation, Genstat usually works with its levels. The exception is when the factor occurs as the first operand of the operators `.IN.` or `.NI.` and the second operand is a text; the factor labels are then used instead. A factor can also receive the results of a calculation; an error is reported if any of the resulting values is not one of the levels of the factor. Two functions are provided especially for factors: the Number of levels function (`NLEVELS`) in the Other class provides the number of levels of a factor, and the Convert factor function (`NEWLEVELS`) forms a variate from the factor supplied by its first argument using the variate supplied by its second argument to define (new) values for the levels.

The button at the top left-hand side of the Calculate menu (alongside the box containing the calculation) opens the Calculation History menu (Figure 2.57). This keeps a record of the calculations that you have done. You can reuse a calculation by clicking on the Replace and Insert buttons at the foot of the menu.

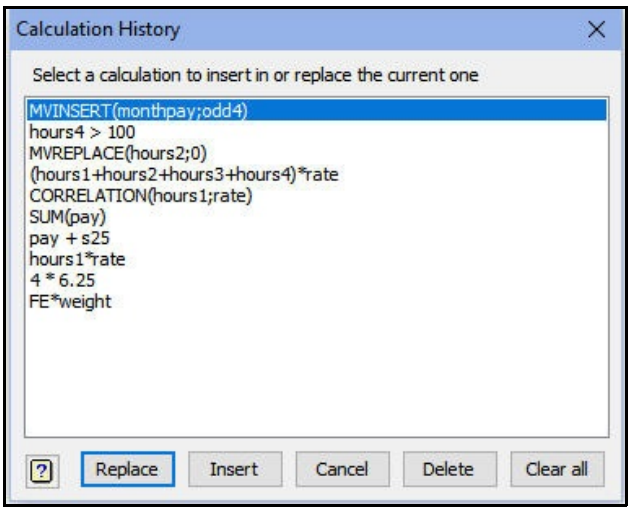


Figure 2.57

2.13 Practical

Details are given below of numbers of personal computers sold by a shop in the months of 2001 and the prices charged; the data are available in the spreadsheet file `Computer.gsh`. Open the spreadsheet using the File menu. Calculate the amount received from PC sales in each month and display them within the current spreadsheet. The price for January has been entered incorrectly and should have been 1099. Change this value and recalculate the column for the amount received. Calculate the total received over the whole year.

```
"month number  price"
January    12     999
February    8    1150
March      21    1150
April      18    1250
May         7    1250
June        5    1250
```

```
July         6    1250
August      18    1099
September    5    1250
October     17    1250
November    13    1250
December    31    1150
```

2.14 Other facilities

This chapter gives only a brief account of the facilities in the Genstat spreadsheet. It can also be used for data input and verification. It can import data from databases and it provides a wide variety of types of data manipulation. Further details are in the *Guide to the Genstat Spreadsheet*, which can be accessed by clicking on the Spreadsheet sub-option of the Genstat Guides option of the Help menu on the menu bar (see Figure 1.6).

2.15 Commands for data input, calculations and display

The menu options illustrated above perform actions that can also be carried out using commands in an Edit window, as outlined in Section 1.3. You may prefer not to use this level of control of the Genstat system, and restrict yourself to what can be done by the menus, but the commands are in fact much more powerful. We outline them here so that you can find out what is available in case you need it. We shall continue to do this in an optional final section of each of the remaining chapters in this book.

The READ directive, described at the end of this section, provides very general facilities for the input of data. However, when the file is in a simple layout (like those handled by the Data menu), it is simpler to use the FILEREAD procedure. The FILEREAD procedure can automatically determine the type of data being read as either numbers or text, and then set up factors in the same way as the Data menu. For example, the data in Bacteria.dat can be input by the following single command:

```
FILEREAD [NAME='Bacteria.dat'] counts,crop; FGROUPS=no,yes
```

Similarly, the alternative data-recording style used in Bacteri2.dat can be handled with the following command (using the continuation character (\) to continue the command onto a second line):

```
FILEREAD [NAME='Bacteri2.dat'; IMETHOD=read; \
  MISSING='-'; SEPARATOR=','] FGROUPS=no,yes
```

You need to be careful with using backslashes when specifying names of directories on a PC. If you need to specify the directory name, you should either duplicate each backslash character, or use the forward slash (/) instead. For example, if the data file was in directory C:\Program Files\Gen21Ed\Data, you could put either

```
Name='C:\\Program Files\\Gen21Ed\\Data\\Bacteria.DAT'
```

or

```
NAME='C:/Program Files/Gen21Ed/Data/Bacteria.dat'
```

The SPLOAD directive can be used to access data within a Genstat spreadsheet. For example to read the data from the file sulphur.gsh you would use the following command:

```
SPLOAD 'C:/Program Files/Gen21Ed/Data/Sulphur.gsh'
```

By default, Genstat will display a list of the data structures within the spreadsheet file. However, you can suppress this summary using the PRINT option as follows:

```
SPLOAD [PRINT=*] 'C:/Program Files/Gen21Ed/Data/Sulphur.gsh'
```

The IMPORT procedure provides a way of importing data from foreign file formats such as Excel, Quattro or dBase etc... The basic use of the IMPORT procedure is the same for most file formats with the exception of Excel and Quattro, which use two additional parameters. The following example shows how to read data from the Genstat Data worksheet from within the Excel file Bacteria.xls.

```
IMPORT 'C:/Program Files/Gen21Ed/Data/Bacteria.xls'; \
  SHEET='Genstat Data'
```

There are two ways that a range of data can be read from an Excel. The first way is to specify the worksheet and the range. For example, the following command will read the range A3:B13 from the Bacteria Counts worksheet.

```
IMPORT 'C:/Program Files/Gen21Ed/Data/Bacteria.xls'; \
  SHEET='Bacteria Counts'; CELL='A3:B13'
```

The second way to read in a range, is to specify a named range in the SHEET option. For example to read in a named range called 'Named_Range' you would use the following:

```
IMPORT 'C:/Program Files/Gen21Ed/Data/Bacteria.xls'; \
  SHEET='Named_Range'
```

You can use the PRINT directive introduced in Chapter 1 to store data in an ASCII file. This directive has a CHANNEL option, so you can open a file for output and instruct PRINT to write data into it with a prescribed format. There are also many options to control details of layout; full details are in the on-line help. For example, the following commands would write a new file containing the same data as within Bacteria.dat, but without the initial comment, and then close the file in order that it could be used by another application:

```
OPEN 'Bacteria.new'; CHANNEL=2; FILETYPE=output
PRINT [CHANNEL=2; IPRINT=*] counts,crop; \
  FIELDWIDTH=3,7; DECIMALS=0
CLOSE 2; FILETYPE=output
```

You can also use the PRINT directive to display the contents of data structures to the Output Window. The following command shows how to display the count and crop structures in the output window.

```
PRINT counts,crop; FIELDWIDTH=3,7; DECIMALS=0
```

You can display data structures in a spreadsheet or add structures to an existing spreadsheet using the FSPREADSHEET procedure. The following command will display the bacteria data in a spreadsheet:

```
FSPREADSHEET count,crop
```

To add data structures you should use the SHEET option. The number provided by SHEET is the position of the spreadsheet in the list of currently open spreadsheets. Thus SHEET=1 will add or update data in the first spreadsheet in the window list, SHEET=2 the second etc. Setting SHEET=0 will cause Genstat to update the first sheet with matching structures (i.e. for a variate this will be a *Vector Spreadsheet* with the same number of rows). The Genstat interface uses internal pointers to the spreadsheet structures which

appear as large integers, and these should not be re-used in your code. For example, the following commands will load count into a spreadsheet and then add crop into this spreadsheet:

```
FSPREADSHEET counts
FSPREADSHEET [SHEET=0] crop
```

The procedure FSPREADSHEET can also be used to save data structures to a Genstat spreadsheet using its OUTFILE option. For example, the following shows how to save the bacteria data to a Genstat spreadsheet:

```
FSPREADSHEET [OUTFILE='MyData.gsh'] counts,crop
```

To save Genstat data structures to a foreign file you can use the EXPORT procedure. The file that the data is saved into depends on the file extension used. For example, if you use the file extension .xls the data will be saved into an Excel file. The file types that are currently supported by Genstat include Excel (.xls), Quattro (.wq1), dBase (.dbf), Splus (.sdd), Gauss (.fmt), MatLab (.mat), Instat (.wor) and comma delimited text files (.csv). To export data to an Excel file you would use the following command:

```
EXPORT ['MyData.xls'] counts,crop
```

If you are using EXPORT in interactive mode then a prompt will appear if you try to overwrite an existing file. You can avoid the prompt for overwriting an existing file with the following command:

```
EXPORT ['MyData.xls'; METHOD=overwrite] count,crop
```

The Calculations menu uses the Genstat CALCULATE directive. This has the form:

```
CALCULATE expression
```

where the *expression* specifies the calculation to be performed, and where the results are to be stored. The expression first indicates the structure (or structures) to store the results (as provided by the Save Result In box of the Calculations menu). There is then an assignment operator =, and then details of the calculation to be done (as listed in the box at the top of the Calculations menu). The CALCULATE statements for the examples discussed in Section 2.12 are reproduced below.

```
CALCULATE s25=6*4.25
CALCULATE pay=hours1*rate
CALCULATE pay1=pay+s25
CALCULATE totalpay=SUM(pay1)
CALCULATE paycorr=CORRELATION(hours1;rate)
CALCULATE monthpay=(hours1+hours2+hours3+hours4)*rate
CALCULATE hours2=mvreplace(hours2; 0)
CALCULATE hours3=mvreplace(hours3; 0)
CALCULATE hours4=mvreplace(hours4; 0)
CALCULATE odd4=hours4>100
CALCULATE monthpay=MVINSERT(monthpay;odd4)
```

Finally, we mention briefly some useful housekeeping commands. To list data structures in the session, you can use the LIST directive. If you type

```
LIST
```

you will be given a list of all the current identified structures.

The DELETE directive can be used to throw away data structures once you have finished with them. By default, it throws away only their values, so you cannot re-use the identifier for another type of structure. However, if you set the REDEFINE option, you can

then re-use the identifier in any way you wish. For example, if we no longer wanted the variate counts, but wanted to set up a scalar called count instead, we could use the commands

```
DELETE [REDEFINE=yes] counts
SCALAR counts
```

If you wish to copy a structure, you can use the DUPLICATE directive. For example, the command

```
DUPLICATE OLD=crop; NEW=newcrop
```

will set up a new factor called newcrop, containing the same values as in crop. If you then delete the original structure:

```
DELETE [REDEFINE=yes] crop
```

this is equivalent to renaming the structure.

Finally we describe the READ directive, which allows you to input data values into any Genstat data structure. In the simplest use of READ, you need specify only the identifiers of the structures to be read. Genstat will then expect you to provide the data values in free format on the next input lines, and to type a colon (:) at the end of the data.

READ has a PRINT option with settings:

summary	to print a summary of the data
data	to print a copy of the input lines
errors	to print a detailed report on any errors in the data

By default PRINT=summary,errors but we include the setting data in all the examples below, so that you can see what is being read. We have also requested Genstat to echo the lines containing the commands. With Genstat for Windows this is requested from the Options menu; with other implementations it happens by default (and can be controlled by the INPRINT option of the SET directive).

All the examples in this section show READ being used in an example program (called Read.gen), which has been executed in batch (by opening it in an input window and then selecting Submit Window from the Run menu on the menu bar, as explained in Section 1.5). READ can also be used in a window that has been set to be interactive (by clicking on the Interactive line in the Tools menu on the menu bar; see Section 1.7). Genstat then expects you to type the data onto the screen. It uses a special prompt string to indicate the unit of the structure whose value is to be read next, and it terminates automatically when enough data values have been supplied (see *Guide to the Genstat Command Language*, Part 1, Section 3.1.2).

We have also checked the Echo Commands box on the Audit Trail tab of the Options menu (Figure 1.36), so that the commands are echoed in the Output window.

```
2  VARIATE [NVALUES=8] rain
3  TEXT [NVALUES=8] day
4  TEXT [VALUES=no,yes] sunlabel
5  FACTOR [NVALUES=8; LABELS=sunlabel] sunshine
```

```
 6  READ [PRINT=data,summary,errors] rain
 7    0  5  14  2.3E1  3
 8  0  2       8 :
```

Identifier	Minimum	Mean	Maximum	Values	Missing
rain	0.0000	6.875	23.00	8	0

```
 9  READ [PRINT=data,summary,errors] day
10  'Last Sunday' Monday   Tuesday  Wednesday     Thursday Friday
11  Saturday   Sunday  :
```

Identifier	Minimum	Mean	Maximum	Values	Missing
day				8	0

```
12  READ [PRINT=data,summary,errors]sunshine; FREPRESENTATION=labels
13    yes  yes     no  no   no       yes  yes   no :
```

Identifier	Values	Missing	Levels
sunshine	8	0	2

The first section of the output reads data into a variate `rain`, a text `day`, and a factor `sunshine`. As you can see, in *free format*, numbers can be given in either ordinary or scientific format (line 7), and they can be arranged in any way you like provided there is at least one space, newline, or tab character between each one. Similar freedom is available for the strings for texts like `day`, but notice in line 10 that the string `Last Sunday` needs to be placed between quotes (`'`). By default, `READ` expects the values of a factor to be represented by its *levels*, but in line 12 we have set the parameter `FREPRESENTATION=labels` to indicate that we shall use the *labels* `no` and `yes` for `sunshine`, instead of the levels 1 and 2.

Next, we print the values of the structure so that you can see what has been read.

```
14  PRINT day,rain,sunshine
```

day	rain	sunshine
Last Sunday	0.000	yes
Monday	5.000	yes
Tuesday	14.000	no
Wednesday	23.000	no
Thursday	3.000	no
Friday	0.000	yes
Saturday	2.000	yes
Sunday	8.000	no

Several structures can be read using a single `READ` statement. Genstat assumes that the values will be read in *parallel* (or unit by unit), and therefore that the structures will all have the same dimensions. This is illustrated in line 15, where we read `rain`, `day` and `sunshine` again, in parallel.

```
15  READ [PRINT=data,summary,errors] day,rain
16  'Last Sunday'  0   Monday  5  Tuesday  14  Wednesday 23 Thursday
17     3   Friday   0     Saturday   2    Sunday   8  :
```

Identifier	Minimum	Mean	Maximum	Values	Missing
day				8	0
rain	0.0000	6.875	23.00	8	0

Structures with different dimensions can be read in *series*, by setting option `SERIAL=yes`. Now the structures are read in turn, and each set of data values has its own terminating colon (`:`), as shown in line 18 of the next section of output.

```
18  READ [PRINT=data,summary,errors; SERIAL=yes] day,rain
19  'Last Sunday' Monday   Tuesday  Wednesday     Thursday Friday
20  Saturday   Sunday  :
21   0  5  14  2.3E1  3
22  0  2       8 :
```

Identifier	Minimum	Mean	Maximum	Values	Missing
day				8	0
rain	0.0000	6.875	23.00	8	0

If a structure whose values are to be read has not already been declared, Genstat will define it automatically as a variate. Likewise, if the length of a *vector* (a variate, text or factor) is undefined, this too will be set automatically. `READ` first checks whether the vector is being read in parallel with other vectors whose lengths have been defined, then it looks to see if a default length has been defined for vectors using the `UNITS` directive. If neither of these is available to define the length, it is set to the number of data values that are provided in the input. For example, in line 23 below, `temp` is defined to be a variate of length five.

If you have declared a structure to be a factor (see Section 2.1), but have not yet defined its levels or labels, `READ` can define these for you too: levels only if `FREPRESENTATION=levels`, or labels (and levels as integers 1, 2...) if `FREPRESENTATION=labels`.

Lengths of vectors can also be *redefined* according to the number of data values that are read, by setting option `SETNVALUES=yes`. This is used in line 25 to redefine the lengths of `day` and `rain` also to be five.

```
23  READ [PRINT=data,summary,errors] temp
24  15.5  12  10.5  18  21 :
```

Identifier	Minimum	Mean	Maximum	Values	Missing
temp	10.50	15.40	21.00	5	0

```
25  READ [PRINT=data,summary,errors; SETNVALUES=yes] day,rain
26  Monday  5  Tuesday  14  Wednesday  23  Thursday  3   Friday  0 :
```

```
Identifier      Minimum         Mean      Maximum       Values      Missing
       day                                                   5            0
      rain       0.0000        9.000        23.00            5            0
```

```
27  PRINT day,rain,temp; DECIMALS=0,0,1

       day         rain         temp
    Monday            5         15.5
   Tuesday           14         12.0
 Wednesday           23         10.5
  Thursday            3         18.0
    Friday            0         21.0
```

For factors, you can set option `SETLEVELS=yes`, to get `READ` to set up the levels or labels according to the values that it finds when reading the data. By default it distinguishes between capital and small letters when forming factor labels, but you can set option `CASE=ignored` to ignore the case of letters. Also, by default the levels or labels are sorted into ascending order, but you can set option `LDIRECTION=given` to leave them in the order in which they are found in the data file.

It is often convenient to have the data in a separate file on the computer, particularly when running Genstat interactively. First you need to open the file on a suitable input channel. This can be done using the `OPEN` directive, as explained earlier in this section. In line 28, we open file `Weather.dat` on input channel 2. We can then read data from this file by setting option `CHANNEL` to 2 in the `READ` statement. Notice that the printed input lines have their own line numbering and a two-space indentation.

A file can contain more than one set of data, and Genstat will remember how far it has read through the file so that you read them in turn. Alternatively, you can rewind the file to start again from the beginning by setting `READ` option `REWIND=yes`. At line 30, we use the `CLOSE` directive to close the file. Channel 2 could then be reused, if necessary, for some other file.

```
28  OPEN 'Weather.dat'; CHANNEL=2
29  READ [PRINT=data,errors; CHANNEL=2] day,rain,temp
   1      Monday   5  15.5   Tuesday  14  12.0   Wednesday  23  10.5
   2    Thursday   3  18.0    Friday   0  21.0   :
30  CLOSE 2
```

Other facilities not described here include the ability to read data in fixed format, to skip data values in free format, and to omit or change the use of colon to mark the end of each set of data (see Section 3.1 of the *Guide to the Genstat Command Language*). Data structures can also be stored and retrieved from special *backing-store* files (Section 3.5 of Part 1 of the *Guide to the Genstat Command Language*).

3 Graphics

3.1 The Genstat graphics wizard

It is often helpful to explore the data using graphical displays. These can be used to examine the structure of the data or to display its distribution. We give a brief overview here. A more detailed account is in the *Guide to Genstat Graphics*.

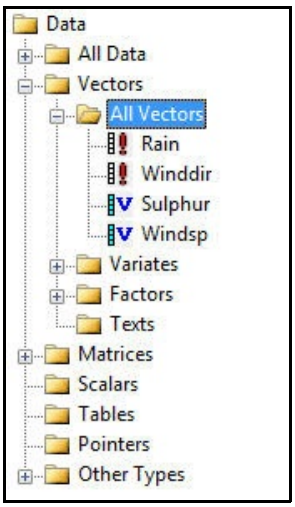


Figure 3.1

Many types of graph can be accessed directly as options of the Graphics menu on the menu bar. Alternative Genstat has a selection tool that can help you choose the right graph. We shall use this to investigate some data collected in 1990 to investigate changing levels of air pollution. The principal measurement is the amount of sulphur in the air each day, but there are also associated measurements: the strength and direction of the wind, and whether or not it rained. The data are available in the file Sulphur.gsh and can be read using the Data file option on the Load menu from the Data menu bar. The Data View pane, shown in Figure 3.1, lists the structures loaded from the file. (You can view the Data View pane by selecting the Data View option on the View menu or the Display option on the Data menu). The data consist of two variates and two factors, each containing 114 values.

You start the process by selecting the Create Graph option from the Graphics menu on the menu bar. The first menu, shown in Figure 3.2, helps you to choose the graph. You indicate the type of data in the top half of the menu, and the corresponding choices are then shown in the bottom half. Here we want to investigate the (variate of) sulphur measurements, and have chosen to investigate their distribution using a histogram. Clicking the OK button opens the menu shown in Figure 3.3.

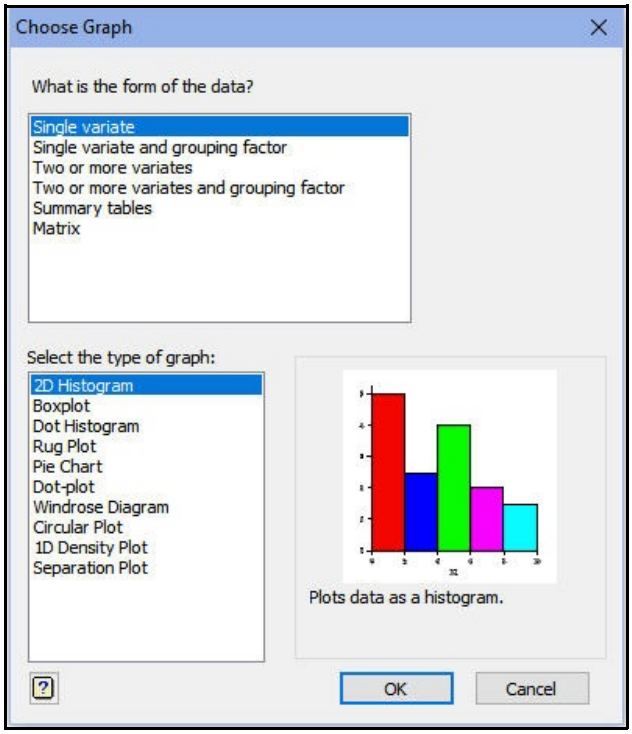


Figure 3.2

The menu opens on the Data tab, where we can enter the data to be plotted. In Figure 3.3, we have selected `Sulphur` in the Available data list, and clicked on the arrow to place it into the Data variates list.

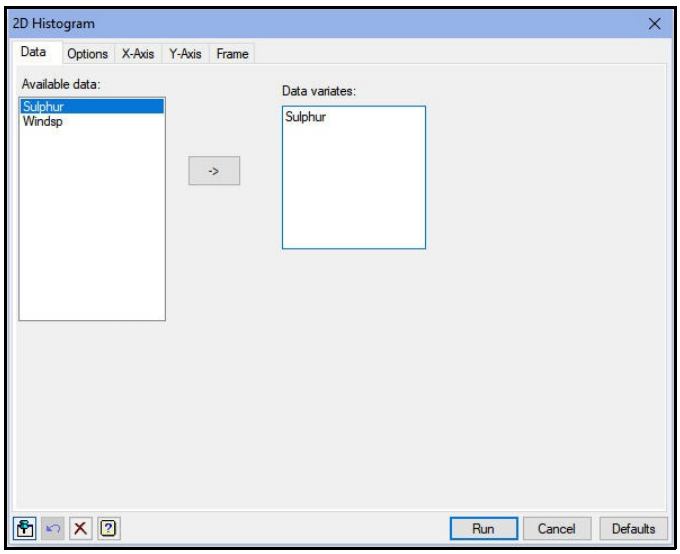


Figure 3.3

We now move to the Options tab, and enter a title, "Sulphur pollution", for the graph. We have selected the Use data values option for specifying the boundaries. This option lets Genstat select the number of groupings and their locations automatically. You can use the Number of groups option to specify a particular number of groups. The groups are then defined by intervals of equal width, spanning the range of values of the variate. Alternatively, if you want to have other intervals, you can define the boundaries explicitly by selecting the Limits option and entering either a variate containing the boundary values, or the list of values themselves.

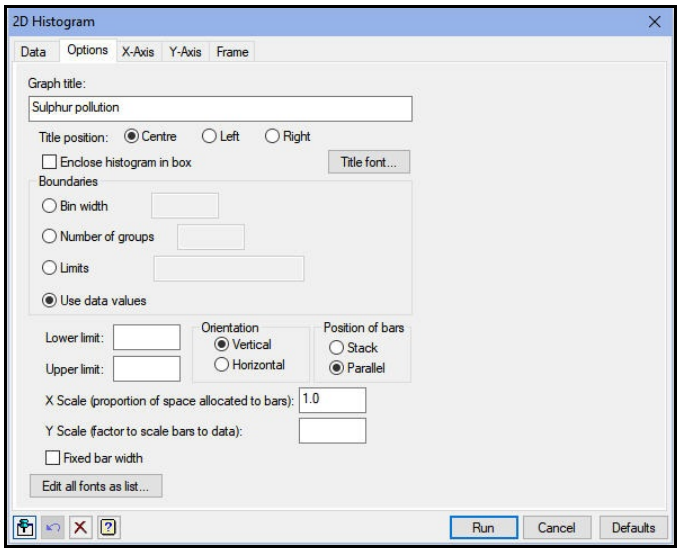


Figure 3.4

Clicking the Run button produces the histogram shown in Figure 3.5. This shows the numbers of observations in successive equal-width categories of the sulphur scale. Clearly sulphur has a skew distribution: there are many days with little or no sulphur in the air, and then decreasing numbers in successive categories with more and more sulphur.

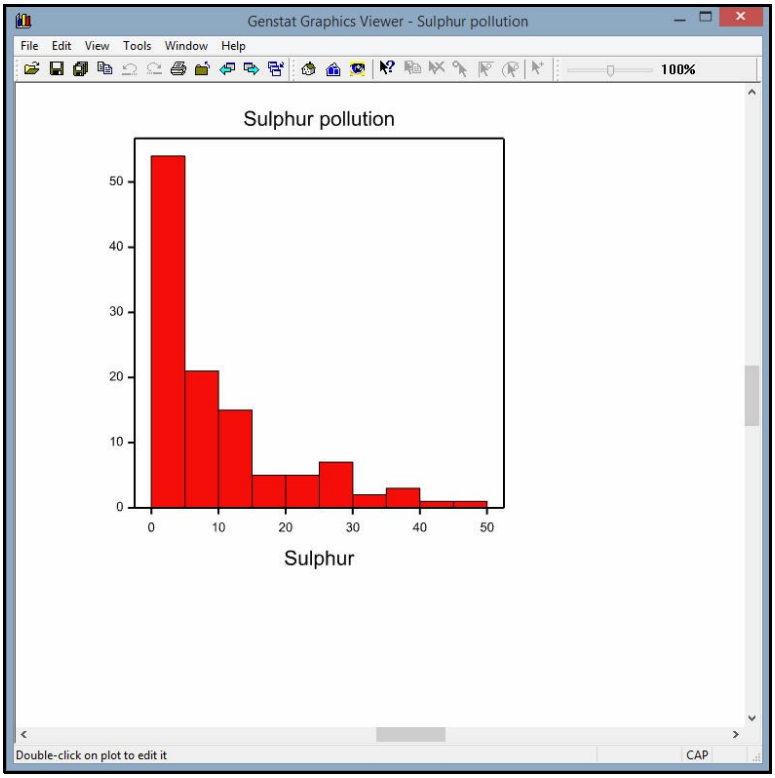


Figure 3.5

Many statistical studies are concerned not with single variables, but with the relationships between several variables. With the pollution it is natural to ask questions like "Is there any effect of wind speed on the sulphur level?"

The most effective way to begin answering a question like this is generally to draw a *scatter plot* or *point plot*. This can be done by selecting Two or more variates option at the top of the Choose Graph menu, and then 2-D Scatter Plot as the type of graph. Clicking on OK opens the menu on the Data tab as shown in Figure 3.6. We wish to plot the sulphur levels against the wind speeds. So we enter `Sulphur` as the Y variate, and `Windsp` as the X variate.

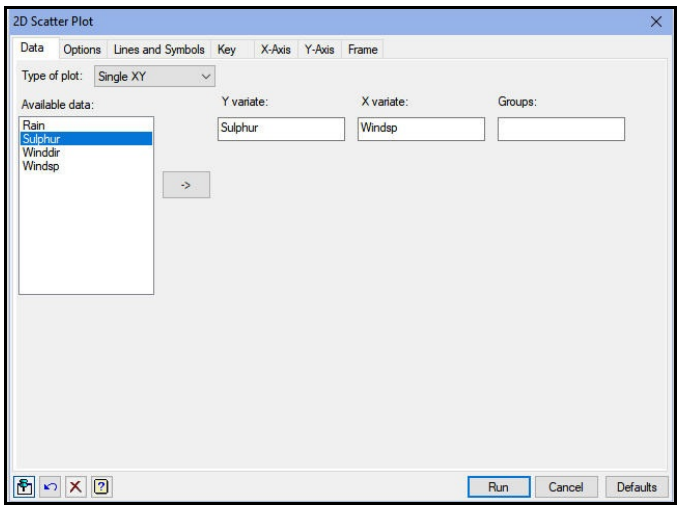


Figure 3.6

The Options tab is shown in Figure 3.7, where we have entered the title into the Graph title box, Clicking on Run produces the graph shown in Figure 3.8. Note, we have set pin, at the bottom-left of the menu, down so that the menu stays open for further plots.

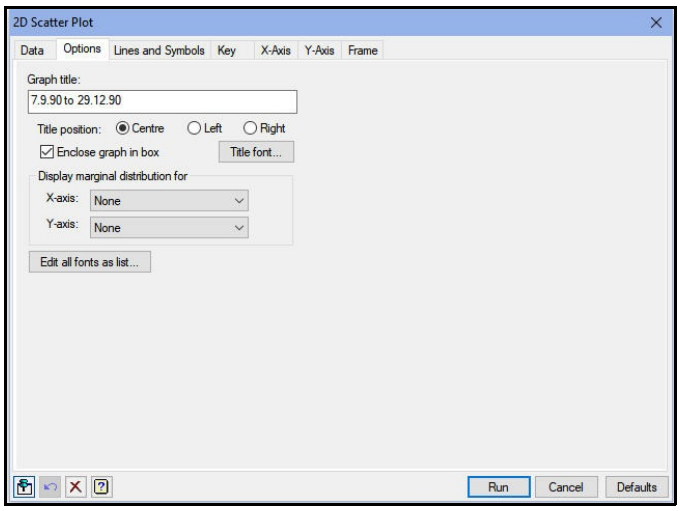


Figure 3.7

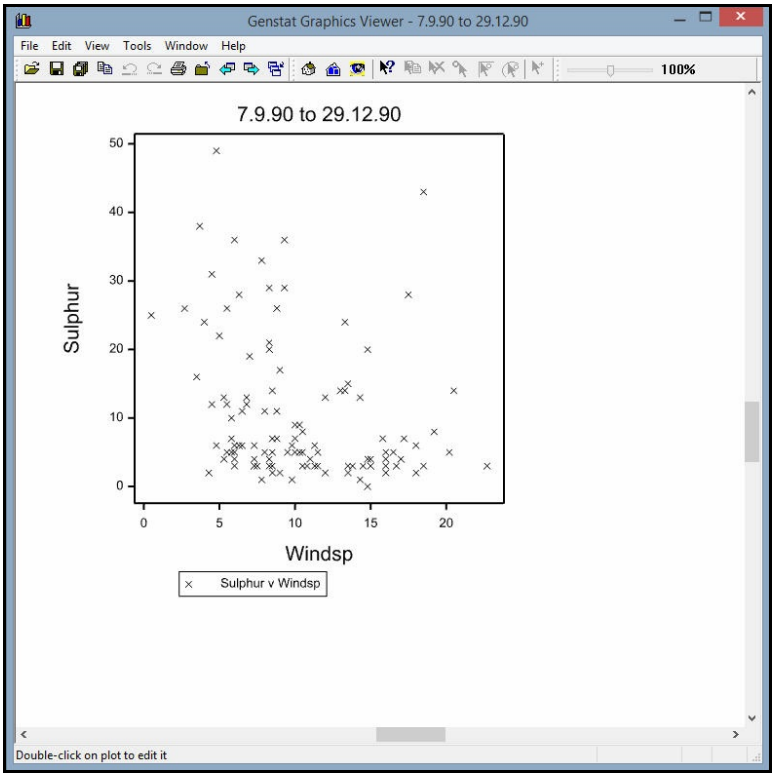


Figure 3.8

You can set the style and colour of the symbols for the scatter plot with the Line and Symbols tab, shown in Figure 3.9. We have selected Plot 1 from the Graph list, and chosen Circle from the symbols. You can chose colours for the circle (in the Colour list box), and for its interior (in the Fill colour list box). Here we have chosen dark blue circles filled with light blue.

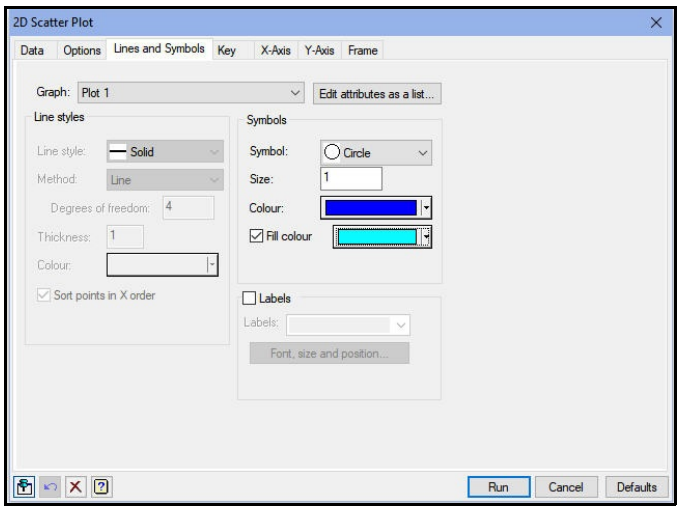


Figure 3.9

Many of the graphics menus contain tabs to allow you to modify the x- and y-axes. These allow you to set attributes such as titles, lower and upper limits, positions of tick marks and their origin positions of the axes i.e., where to draw the x-axis across the y-axis, and vice-versa.

The X-Axis tab is shown in Figure 3.9. To improve the plot, we have selected the Display title option and entered the title as `Wind speed m/s`. We have also entered lower and upper bound values of 0 and 25.

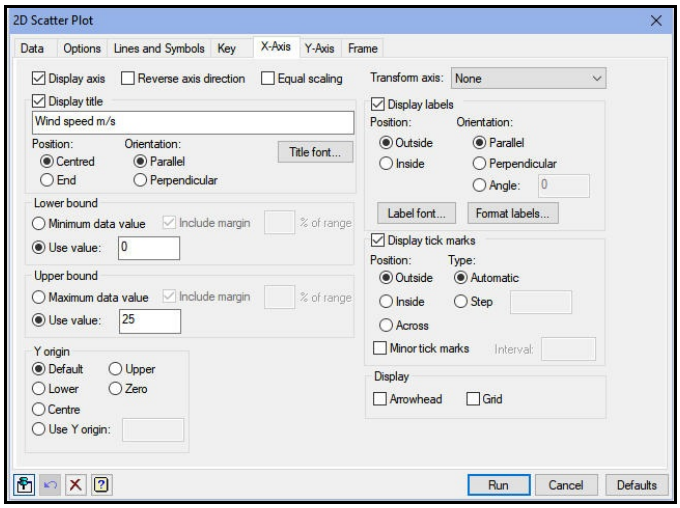


Figure 3.10

We now click on the Y-Axis tab which produces an identical menu to the X-Axis tab. The new menu is shown in Figure 3.11 where we have selected the Display title option and entered the title `Sulphur microg/m~^{3}` into the space provided. The string `~^{3}` is a special typesetting command to make 3 a superscript letter. More details on typesetting can be found in Section 1.4 of the *Guide to Syntax and Data Management*.

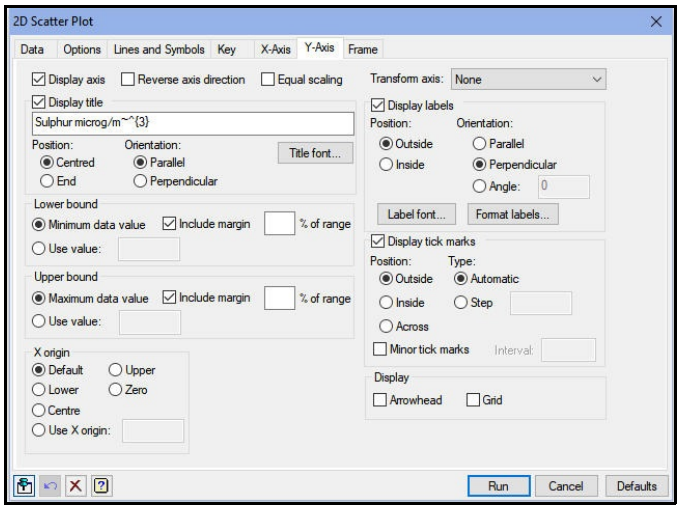


Figure 3.11

The Key tab, shown in Figure 3.12, controls the appearance of the key. Here we have simply unselected the option Display key so that no key is displayed on the graph.

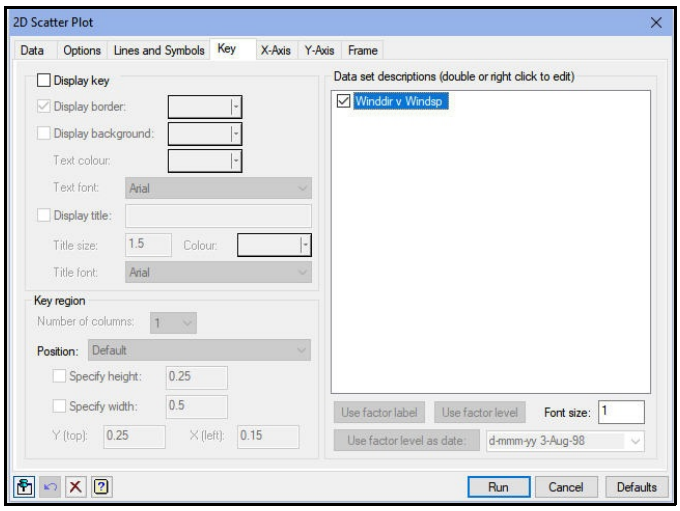


Figure 3.12

Finally, the Frame tab (Figure 3.13) allows you to control the positioning of the plot within the graphics "frame". Each graph is plotted into one of the "windows" in the frame. Here we are plotting into window 1 using its default position. This is defined to use the top 3/4 of the frame. (Window2, which is used if you include a key to the graph, uses the lower 1/4 of the frame.) The Draw radio buttons allow you to put extra graphs onto an existing frame.

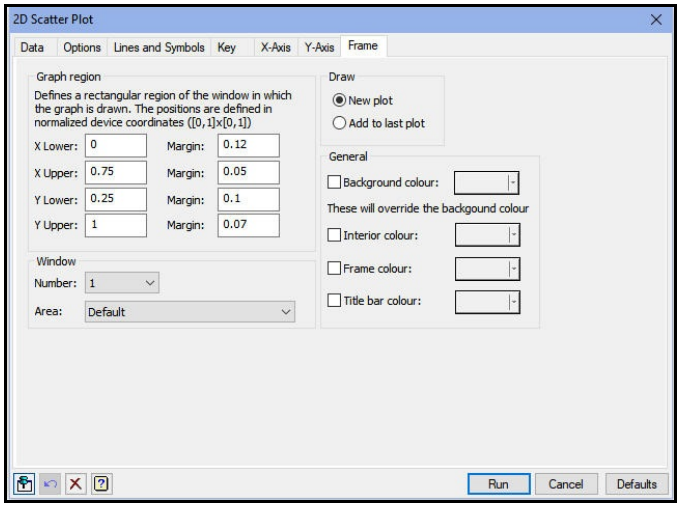


Figure 3.13

Clicking on Run produces the graph shown in Figure 3.14, where the points are now represented by filled circles.

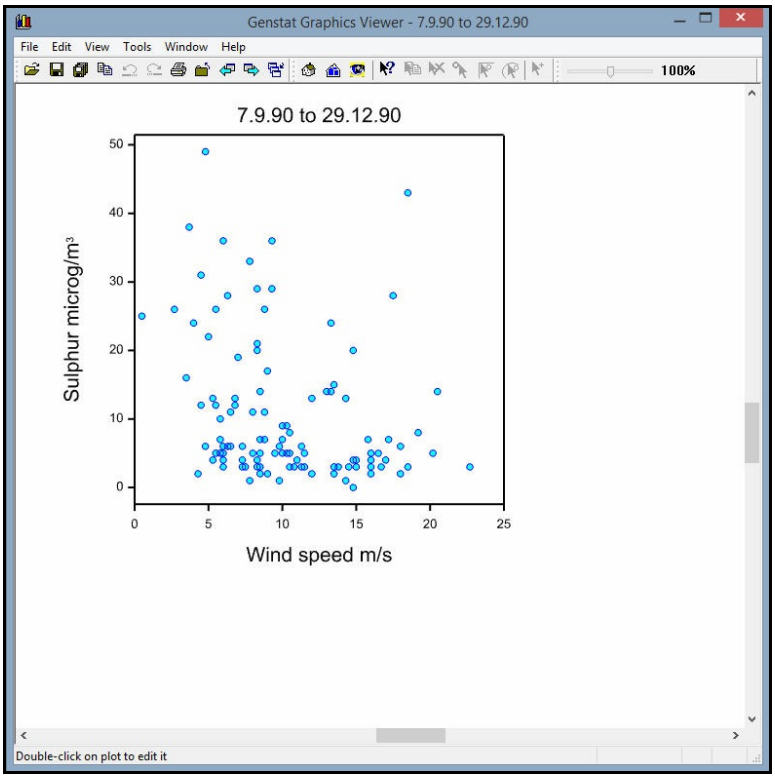


Figure 3.14

3.2 Practical

The number of male deaths from lung cancer per million population, and the average number of cigarettes smoked by men in 1930 in 11 countries is given below and is available in `Smoking.gsh`.

Country	Deaths	Smoking rate
Australia	172	452
Canada	151	508
Denmark	168	379
Finland	353	1113
'Great Britain'	468	1145
Holland	244	468
Iceland	60	226
Norway	95	258
Sweden	116	315
Switzerland	252	540
USA	194	1290

(Data from Tufte, 1983, *The Visual Display of Quantitative Information*. Graphics Press: Cheshire, Connecticut, USA.)

Draw a graph of cancer incidence against smoking rate, providing informative titles. Set axis limits to make sure that both the y-axis and x-axis start at 0. Also, draw the symbols for the points as red circles using the Line and Symbols tab.

3.3 Graphics environments

When you start Genstat an initial graphics environment is set up which contains default settings that are designed to be appropriate for the different types of plots. A graphics environment specifies how graphs are produced controlling aspects such as whether or not boxes are drawn around the plots, the display of the key, and the styles and the colours of outlines for graphs such as histograms or shade plots.

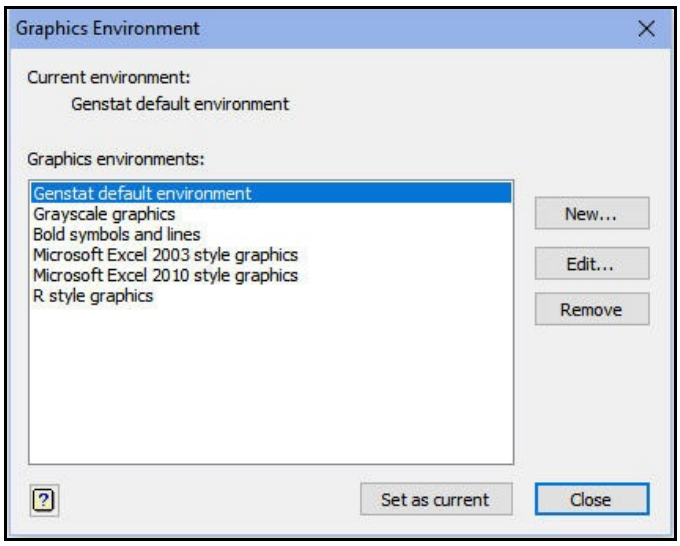


Figure 3.15

Genstat provides a range of environments to suit different tastes and situations. To see the choices, select the Graphics environments option from the Tools menu bar. This opens the menu shown in Figure 3.15. You can highlight another environment in the Graphics environments window, and then click on the Set as current button to use that instead. You can design your own graphics environment, by clicking on the New button. Alternatively, you can change an existing environment by clicking on the Edit button.

3.4 Commands for graphics

The histogram in Figure 3.5 can be produced by the DHISTOGRAM directive as follows

```
DHISTOGRAM sulphur
```

The DHISTOGRAM directive arranges the layout and labelling of the picture itself, but there are many options available if you want to customize the display. For example, you can define your own ranges for the bars by setting the LIMITS option to a suitable variate; this corresponds to the Limits box in the menu in Figure 3.4. Likewise, the NGROUPS option corresponds to the Number of Groups box.

An extension to the edit capabilities available in command mode is that you can change the colour in which the bars are drawn. This is controlled by the Genstat *pen* that is used to draw the bars. For example

```
DHISTOGRAM sulphur; PEN=2
```

would draw the histogram bars in red rather than white, as this is the colour associated by default with pen 2 in Genstat. (We describe later how to change the default attributes of the pens.)

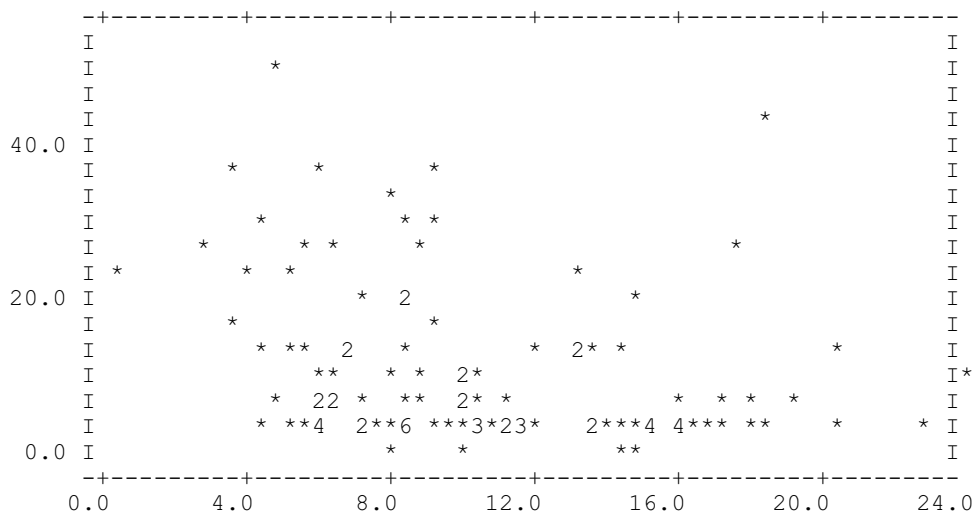
Scatter plots are provided by the DGRAPH directive for high-resolution plots, and by LPGRAPH for line-printer plots. For example, the plot in Figure 3.8 can be produced by the statement

```
DGRAPH sulphur; windsp
```

However,

```
LPGRAPH sulphur; windsp
```

would give the line-printer alternative. The digits in the picture, shown below, represent coincident points in the coarse grid used by LPGRAPH (the digit 9 would indicate nine or more coincident points).



As with the DHISTOGRAM directive, there are many details that can be modified if you are not satisfied with the default settings. For example, DGRAPH also has a PEN parameter that allows you to change attributes such as the colour used to plot the points, so we could set PEN=2 to draw the crosses in red. It can also be used to change the plotting symbol in conjunction with the PEN directive, the PEN parameters of interest here being SYMBOL and CSYMBOL. For example, the command

```
PEN 2; SYMBOL='triangle'; CSYMBOL='green'
```

sets up pen 2 to draw triangles rather than crosses, and display them in green in the PC implementation. You can specify the symbols either by giving the name (as a string), or by number. The available symbols, with their corresponding numbers, are as follows.

1 'Cross'
2 'Circle'
3 'Plus'
4 'Star'
5 'Square'
6 'Diamond'
7 'Triangle'
8 'Nabla'
9 'Asterisk'
10 'Minus'
11 'Heavyminus'

```
12 'Heavyplus'
13 'Heavycross'
14 'Smallcircle'
15 'Tinycircle'
16 'Female'
17 'Male'
18 'Rhombus'
19 'Circlecross'
20 'Circleplus'
21 'Squarecross'
22 'Squareplus'
-1 'Sphere'
-2 'Cone'
-3 'Cylinder'
-4 'Cube'
```

The final four symbols (numbered -1 to -4) are intended for 3-dimensional plots.

Axis titles for high-resolution histograms and scatter plots are specified using the `XAXIS and YAXIS` directives. These provide detailed control of the axes in the picture, for example, the following commands specify titles for both the x-axis (horizontal) and the y-axis (vertical):

```
XAXIS 1; TITLE='Wind speed m/s'
YAXIS 1; TITLE='Sulphur microg/m**3'
```

The digit `1` here refers to the number of the graphical *window* whose axes are to be specified. By default, `DGRAPH` and `DHISTOGRAM` draw their pictures in Window 1, which takes up most of the screen, and their keys in Window 2, which is a narrow rectangle along the bottom of the screen. You can use up to 32 windows, so you can position several pictures on the same screen.

Most graphics commands also have a `TITLE` option to supply a general title for the plot.

High-resolution graphics can be produced on media other than the screen. This is done by selecting a different graphical *device*. The possibilities can be found by the following command:

```
DHELP
```

All aspects of the graphical "environment" in Genstat are given initial settings relevant to each device, but they are re-definable if necessary (see the *Guide to the Genstat Command Language*, Part 1, Section 6.8).

You can save the current graphical environment settings to a file using the following command:

```
DSAVE 'mysettings.gev'; DESCRIPTION='My graphics settings'
```

Once stored the file `mysettings.gev` can be opened to set the current graphics environment to the stored settings.

```
DLOAD 'mysettings.gev'
```

4 Basic statistics

In this chapter we describe how you can produce summary statistics and introduce some analytical statistics. We remind you how to use Genstat to calculate descriptive statistics, such as the mean and median, from a set of observations, and show how to produce summary tables from categorical data. Many statistics are best viewed graphically. In Chapter 2 we described how to draw histograms and scatter plots; in this chapter we show how to draw bar charts and boxplots. We also include here some standard methods for comparing groups of observations: the t-test, corresponding nonparametric tests and χ^2 tests.

4.1 Comparing two samples

Much of analytical statistics is concerned with comparisons, whether of categories of people in medicine or sociology, of animals or plants in biology or agriculture, or of machinery or processes in industry. We consider here only the simplest type of comparison; when there are just two samples of a single measurement. More complicated situations are covered in Chapter 5 onwards.

We shall look at two samples of measurements of sulphur in the air: those taken on rainy days are compared with those taken on dry days. (The data are in `Sulphur.gsh`). The intention is to explore whether there is a difference between the amount of sulphur present in the air on wet and dry days – this is quite likely from a scientific basis, because it is known that rain tends to wash sulphur out of the air. In Chapter 2 we described the importance of first exploring the structure and distribution of your data using graphical methods before performing any statistical tests. It can also be useful to look at numerical summaries.

The data sets that are used in the examples and exercises in this Guide can be all be accessed from within Genstat using the Example Data Sets menu. To open the menu, you click on File on the menu bar, and select the Open Example Data Sets option as shown in Figure 4.1.

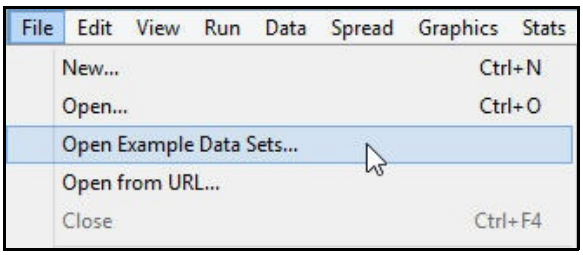


Figure 4.1

In the menu (Figure 4.2) it is easier to find the relevant file if you set the Filter by topic drop-down list to Introduction to Genstat for Windows. We select the file Sulphur.gsh and click on the Open data button.

The summary statistics and graphical displays can obtained from the Summary Statistics menu as described in Chapter 1. To produce the menu you select the Summary Statistics sub-option of the Summary Statistics option of the Statistics menu, as shown in Figure 1.14.

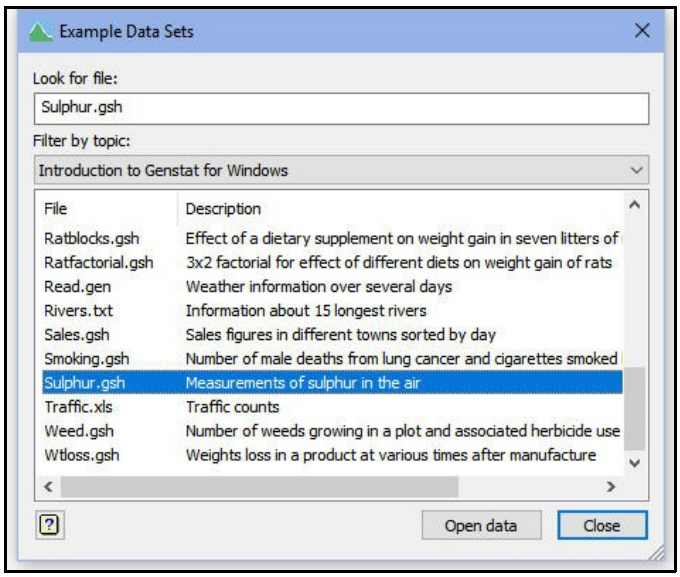


Figure 4.2

Figure 4.3 shows the summary options that we have chosen for this example, and we have again clicked the Boxplot option to display boxplots of the samples. If we enter `Sulphur` into the Variate box, and `Rain` into the Groups box Genstat produces the output below and the boxplots in Figure 4.4.

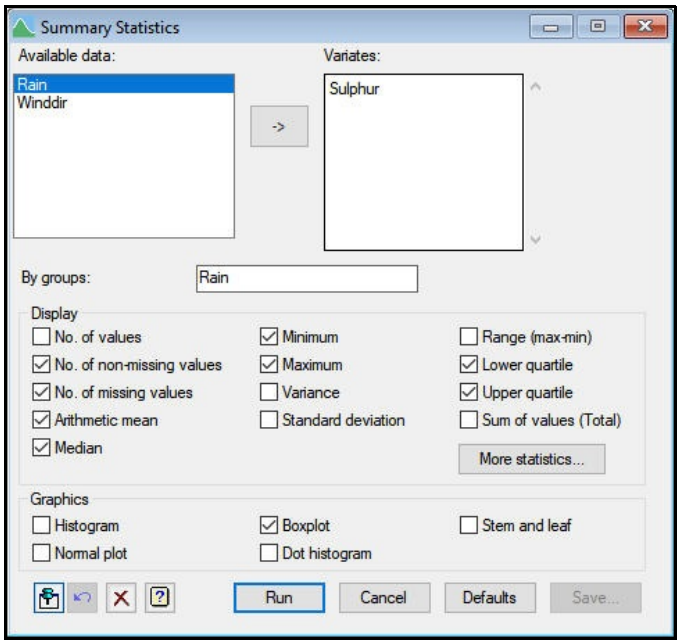


Figure 4.3

Summary statistics for Sulphur: Rain no

Number of observations =	64
Number of missing values =	0
Mean =	12.09
Median =	7
Minimum =	0
Maximum =	49
Lower quartile =	4
Upper quartile =	16.5

Summary statistics for Sulphur: Rain yes

Number of observations =	50
Number of missing values =	0
Mean =	8.38
Median =	5
Minimum =	1
Maximum =	38
Lower quartile =	3
Upper quartile =	10

The numerical summary indicates that there is indeed a higher average sulphur content in the air on dry days. The boxplot shows that the distribution of sulphur values is more squashed towards zero on wet days, although there seem to be no wet days with precisely zero sulphur.

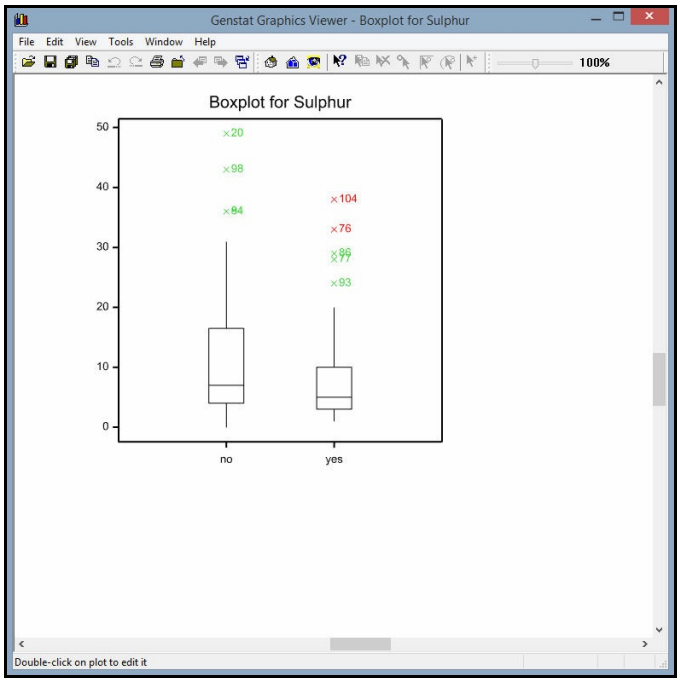


Figure 4.4

We shall now move from descriptive to analytical methods to describe ways of carrying out formal statistical tests on the samples. The first test we shall look at is the most well-known two-sample test, the t-test. This type of test should only be used when the distributions of the groups of observations are reasonably Normal with similar variances. However, the boxplot (Figure 4.4) shows that the distributions are very skew, so the assumption of Normality is not justified. We could, however, try changing the scale of the data by transforming the values to logarithms to satisfy the assumption of Normality.

As described in Section 2.12, Genstat has a general Calculate menu that can be used to transform data. To produce the menu, select the Calculate option from the Data menu. The Calculate menu, as shown in Figure 4.5, contains a box at the top into which you enter the expression i.e., what you want to evaluate in your calculation.

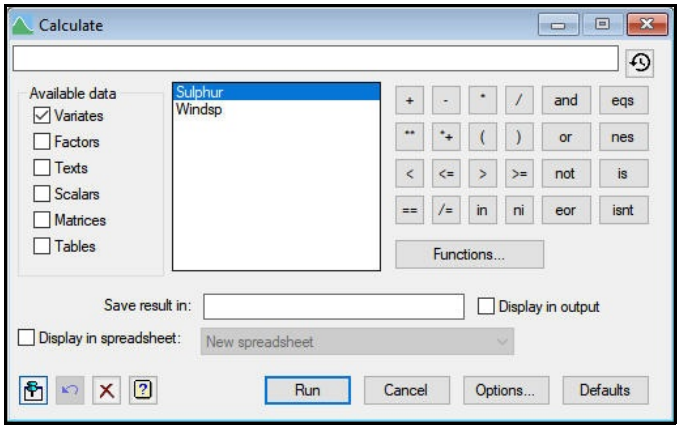


Figure 4.5

To calculate the logarithmic values we need to build an expression using the standard logarithmic function provided by Genstat. This function is available within the Calculate Functions menu which can be found by clicking on the Functions button.

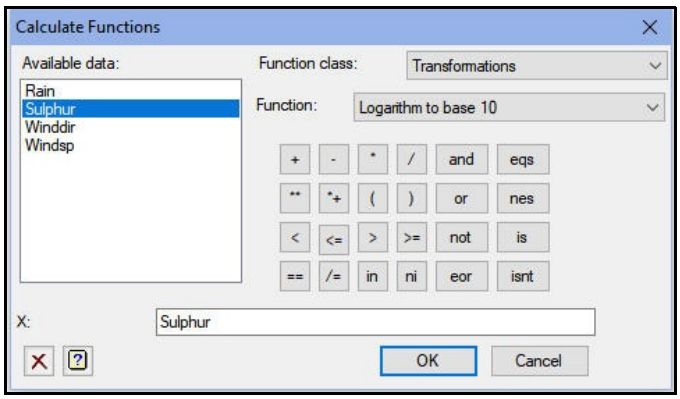


Figure 4.6

In the Calculate Functions menu (see Figure 4.6) the logarithmic function is stored within the Transformations class which contains all the standard logarithmic, trigonometric and statistical transformations as well as absolute value, integer rounding and truncation, differences, shifts and square roots. So, we select Transformations from the Function class drop-down list, and then Logarithm to base 10 from the Function list. This function has a single argument (the numbers to be logged), and you enter this into the X box from the Available data box. When you click OK, the function and its argument are transferred to the main Calculate menu, and inserted into the calculation at the current cursor position. Figure 4.7 shows the expression formed for the calculation, where the name of the function within Genstat is `LOG10`. We shall store the transformed values in `LogSulphur` by entering the name into the Save result in box and clicking Run.

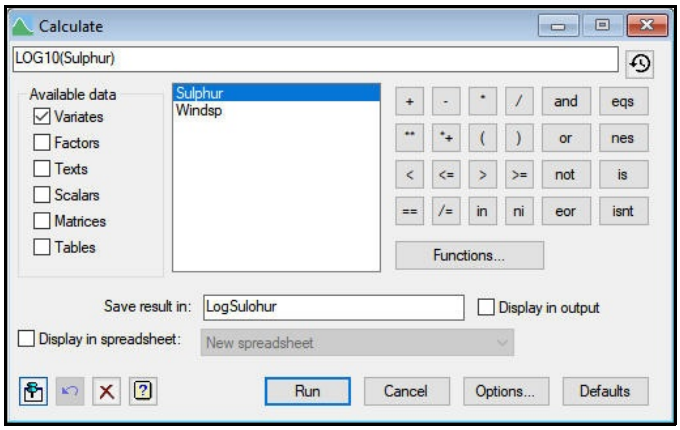


Figure 4.7

Since some of the observations are zero, Genstat gives a warning by producing the dialog shown in Figure 4.8. This has buttons that allow you to move to the description of warning in the Output Window or to open the Event Log.

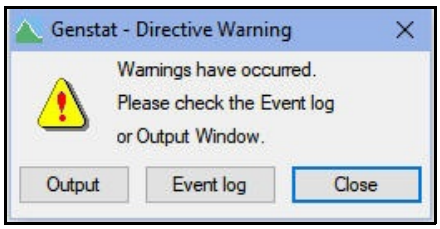


Figure 4.8

The Event Log, shown in Figure 4.9, keeps a record of faults, warnings, and other "events" such as restarting the server or clearing the Output window. A double-click on a line in the log containing a fault or warning takes you to its description within the Output Window. So the Event Log provides a convenient way of reviewing the various activities and mishaps that have taken place during your analysis.

The warning, shown below, tells us that we cannot take logs of zero values. In this situation Genstat will insert a missing value in the variate LogSulphur for any zero values in the variate Sulphur. This should also warn us to be careful in interpreting the analysis, since the logarithmic transformation has to be used with caution when applied to values that may range right down to zero.

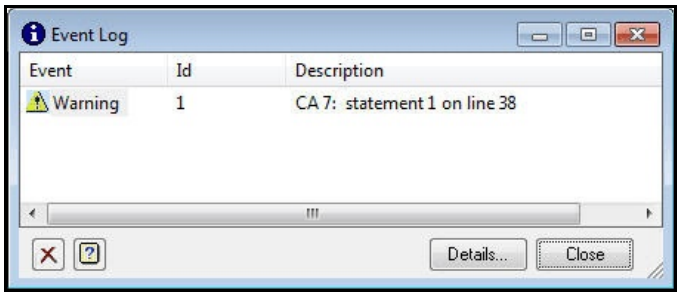


Figure 4.9

Warning 1, code CA 7, statement 1 on line 145

Command: CALCULATE LogSulphur=LOG10(Sulphur)
Invalid value for argument of function.
The first argument of the LOG10 function in unit 1 has the value 0.0000

To evaluate the transformation, we shall examine a boxplot of the transformed values. This time we use the Boxplot menu from the graphics wizard, which offers some additional options not available in the Summary of Variates menu. The menu (Figure 4.10) can be obtained from the Graphics Wizard or, more directly, by selecting the Boxplot option of the Graphics menu.

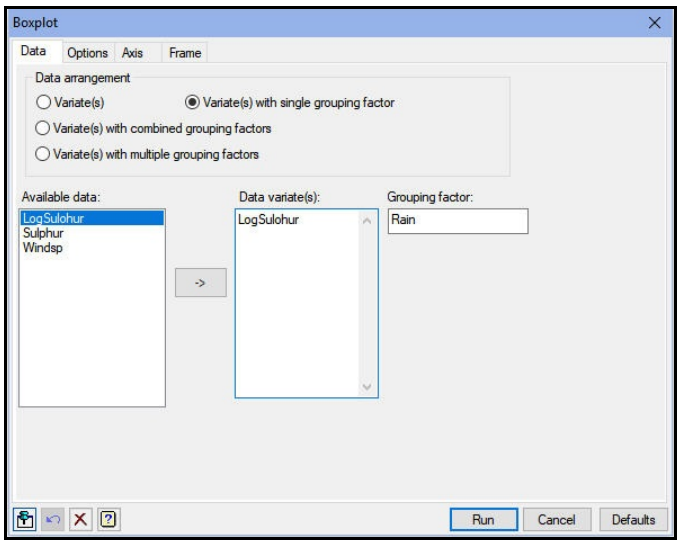


Figure 4.10

The Data tab provides four ways of specifying the data, using the radio buttons in the Data arrangement box. In Figure 4.10 we have indicated that we have a variate together with a factor defining some groupings. The variate is LogSulphur and the factor is Rain, which contains the groupings (yes and no). The menu also has other tabs, like those in the 2D Histogram menu (Figure 3.4) to define titles, axis definitions and so on.

Clicking the Finish button produces the graph, in Figure 4.11. This graph shows that the distributions of the transformed data are much more symmetrical and the variances are more equal. Therefore, we can proceed to carry out a t-test.

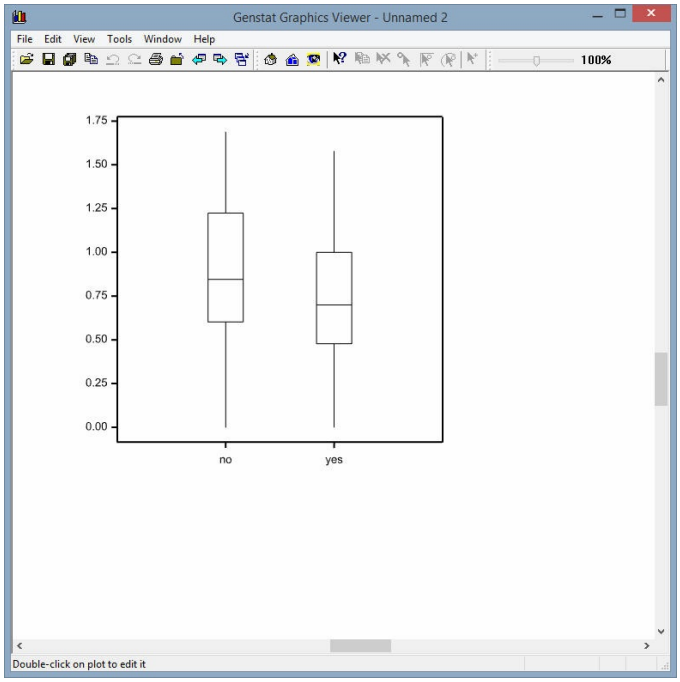


Figure 4.11

To produce a t-test, you first click Stats on the menu bar, select Statistical Tests and click on One and two-sample t-tests. The type of t-test is selected using the Test list box. The possibilities include one-sample, unpaired two-sample and paired two-sample – each with an appropriate selection of boxes and buttons. As the sulphur data consists of two samples of data, we select the two-sample test which produces the menu in the form shown in Figure 4.12.

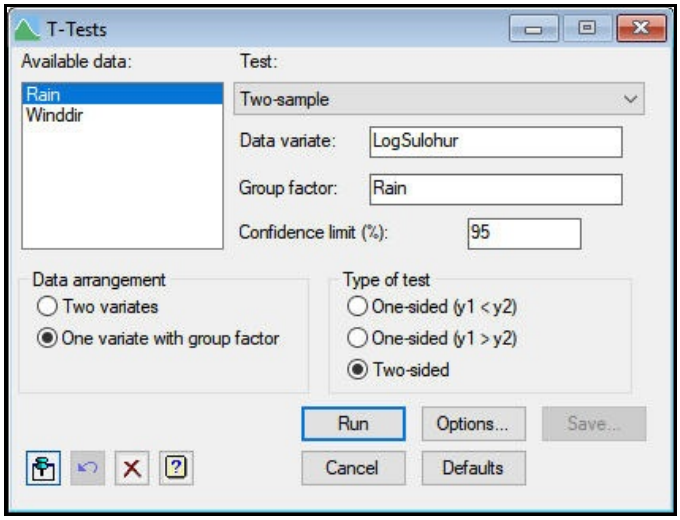


Figure 4.12

The two-sample (unpaired) test provides two ways of specifying the samples, using the radio buttons in the Data arrangement box: either in two separate variates or by specifying a variate together with a grouping factor. In our example we have the variate `LogSulphur` and the factor `Rain`. You can control the form of the test by selecting one of the Type of test options. The possibilities include One-sided (y1 < y2), One-sided (y1 > y2) and Two-sided tests. In this example we select a two-sided test.

Figure 4.13 shows the T-Test Options menu (selected by clicking the Options button), that allows you to control the output produced initially from the analysis and to choose which method to use to estimate the variance. As we described earlier, the t-test assumes that the variances are similar. The reasonableness of this assumption should always be considered in terms of the type of data. A statistical judgement can also be made by comparing the ratio of the two sample variances against 1 using an F-test. If the variances are similar (i.e. the F-test is close to 1), we can use the pooled estimate of variance. Otherwise it may be more sensible to use a separate estimate for each group. Here we have selected the estimation of the variance to be Automatic, so that if there is no evidence that the sample variances are unequal, the pooled estimate will be used in the test; otherwise, separate estimates will be used. The default output is as follows:

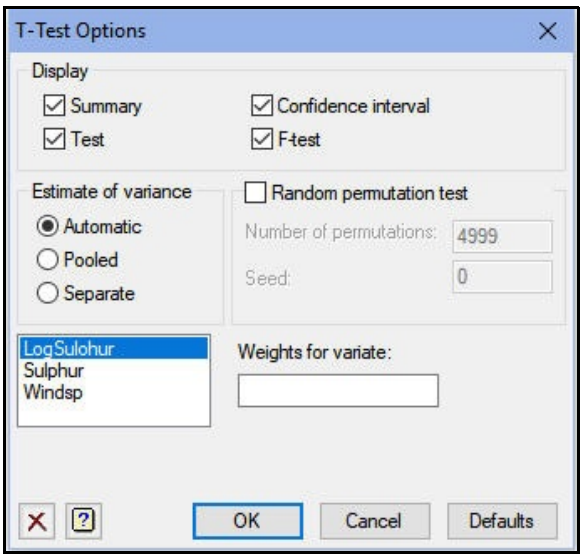


Figure 4.13

Two-sample t-test

Variate: LogSulphur
Group factor: Rain

Test for equality of sample variances

Test statistic F = 1.08 on 62 and 49 d.f.

Probability (under null hypothesis of equal variances) = 0.79

Summary

Sample	Size	Mean	Variance	Standard deviation	Standard error of mean
no	63	0.9216	0.1544	0.3929	0.04951
yes	50	0.7494	0.1431	0.3783	0.05350

Difference of means:	0.17219
Standard error of difference:	0.07321

95% confidence interval for difference in means: (0.02712, 0.3173)

Test of null hypothesis that mean of LogSulphur with Rain = no is equal to mean with Rain = yes

Test statistic t = 2.35 on 111 d.f.

Probability = 0.020

The output contains a summary (mean, variance standard deviation and standard error of mean), test statistic, confidence interval and F-test for checking the assumption of equal variances. We can see that the probability of having equal sample variances is 0.79, so Genstat will use the pooled estimate of variance for the t-test. We can conclude that the logarithms of the sulphur values are significantly smaller on rainy days than dry days, and so infer that the sulphur values are themselves smaller. The difference between the two means is 0.1722, so we can also say that on average the sulphur level is about 49% higher on dry days (the antilog of 0.1722 is 1.487, or about 149%).

Instead of transforming the sulphur data, we might consider an alternative nonparametric test that does not make any strong assumptions about the actual form of the distributions. One possibility is a nonparametric test called the *Mann-Whitney U test*. This involves calculating a test statistic from the relative orders of the observations in the two categories. The statistic should be small if there is little difference between the samples; the test is made by calculating the probability of getting a value at least as extreme as that observed if there were indeed no difference.

To produce a Mann-Whitney test you select Statistical Tests from the Stats menu and click on Two-sample nonparametric tests. In this menu select the Mann-Whitney U test from the drop-down list box of tests and select the type of data arrangement to be One variate with group factor; see Figure 4.14. We select the Data variate to be `Sulphur` and the Group factor to be `Rain` and click on Run.

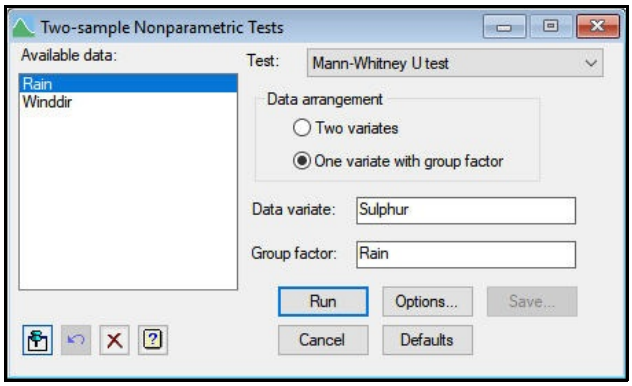


Figure 4.14

Mann-Whitney U (Wilcoxon rank-sum) test

Variate: Sulphur
Group factor: Rain

Value of U: 1226.5 (first sample has higher rank sum).

Exact probability (adjusted for ties): 0.033
(under null hypothesis that group no is equal to group yes).

Sample sizes: 64, 50.

The output here shows that the first level of the factor (group "no") has higher values than the second level (group "yes"), and has the exact probability of 0.033 if there were no difference between the distributions in the two categories. With a probability as small as this we conclude that there is evidence of more sulphur in the air on dry days, as we found before.

4.2 Practical

Twelve people were tested to investigate the relationship between reflex blink rate and the difficulty of performing visual motor tasks. Here are their recorded blink rates during a simple and a difficult task:

Subject	Simple	Difficult
1	14.0	5.0
2	19.5	6.6
3	8.2	1.9
4	8.5	1.5
5	12.1	1.1
6	8.0	2.5
7	8.2	0.6
8	10.1	0.5
9	5.5	0.5
10	10.1	3.1
11	7.2	2.1
12	5.6	1.6

These results are stored in two variates, one for each task, in the spreadsheet file `Blink.gsh`. Display the two distributions side-by-side using boxplots. Form and display the means and standard deviations for each task. Carry out a test of the hypothesis that subjects have the same blink rate in the two tests. You need to use a *paired test* because the same subjects were used for both tasks: if you decide to avoid strong assumptions, use the Wilcoxon test procedure rather than the Mann-Whitney test.

4.3 Summarizing categorical data

When the data values are categorical rather than continuous, different kinds of summary are needed. For example, the information about wind direction cannot be summarized easily with means or quartiles. Instead we might want to count the numbers of observations in each category.

We can do this by selecting the Frequency Tables sub-option of the Summary Statistics option of the Stats menu on the menu bar. This opens the Frequency Tables menu (Figure 4.15), which tabulates frequency counts for grouped data. Here we want to print a table showing how often each level of the factor `Winddir` occurs, so we enter `Winddir` into the Groups box. We shall also save the counts so that we can plot them later. So we check the Save frequencies box, and type in `Ndirdays` as the identifier of the Genstat *table* structure in which we want to save them. Clicking on Run produces the output below.

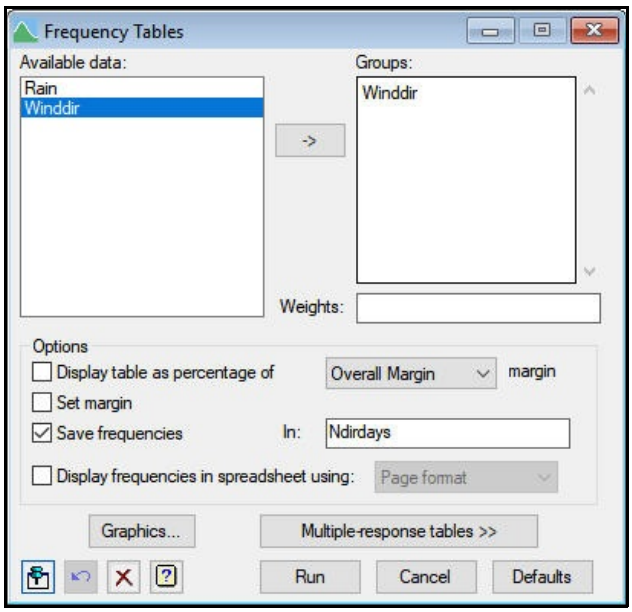


Figure 4.15

```
Ndirdays
   Winddir
         E          8
         N          8
        NE         11
        NW         11
         S         15
        SE         11
        SW         28
         W         21

Unknown cell
                Ndirdays          1
```

Here we have a *one-way* table – that is, it is classified by only one factor (Winddir) – but Genstat allows two-way tables, as you will see in Section 4.6. In fact tables involving up to nine factors can be produced. The *Unknown cell* summarizes the units where the factor has a missing value. So, the output shows that there is one day when the wind direction was unknown.

You can also display the counts in a *bar chart*. This is similar to a histogram, but the categories are not necessarily on a continuous scale, nor even in a particular order. To draw a bar chart, we select the Bar chart option of the Graphics menu. Figure 4.16 shows the Bar Chart menu that then appears. We already have the table of counts, `Ndirdays` to plot. So we select Summary table(s) as the Data arrangement, and enter `Ndirdays` as the Summary table.

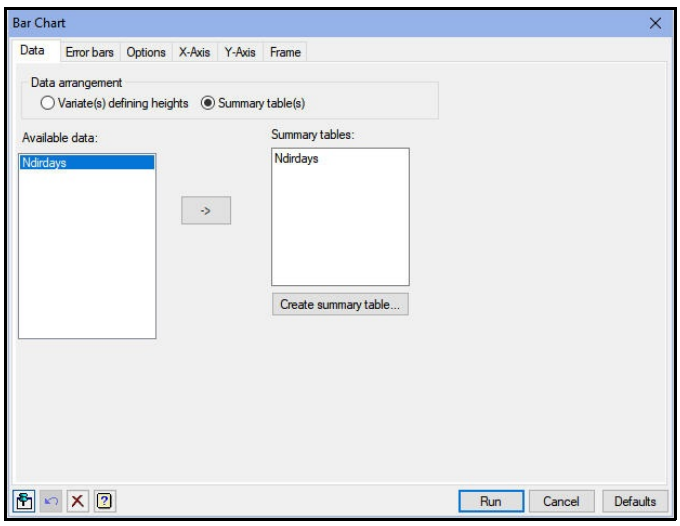


Figure 4.16

Clicking on Run then plots the bar chart, as shown in Figure 4.17. The bar chart differs from the histogram in the style of labelling, and the fact that the bars are slightly separated. Here the picture shows clearly how the wind direction varies, emphasizing that the prevailing wind is from the South-west.

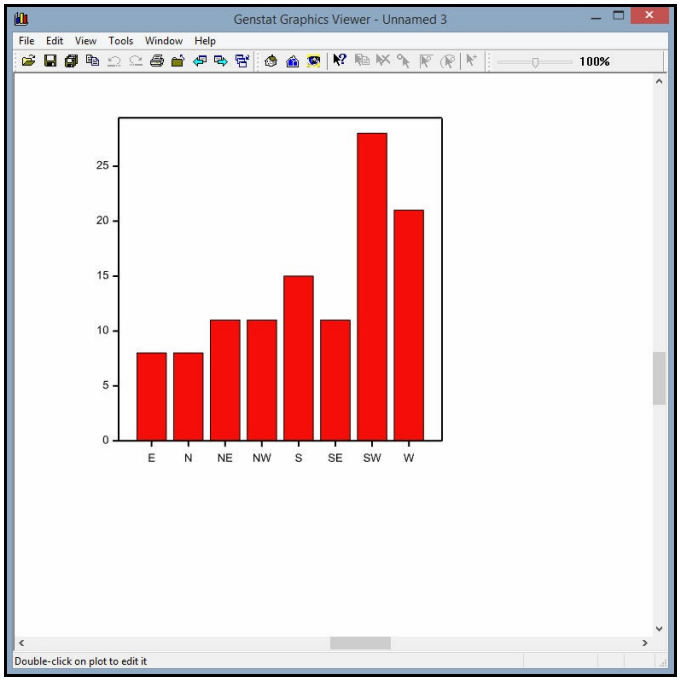


Figure 4.17

4.4 Summarizing data by groups

You can form tables of summary statistics for grouped data using the Summary Tables menu (Figure 4.18). To produce the menu you select Summary Tables sub-option of the Summary Statistics option of the Stats menu. The most popular summary statistics are available in the Display box. You can also click on the More button to open a subsidiary menu, More summary statistics (Figure 4.19), to access some additional statistics.

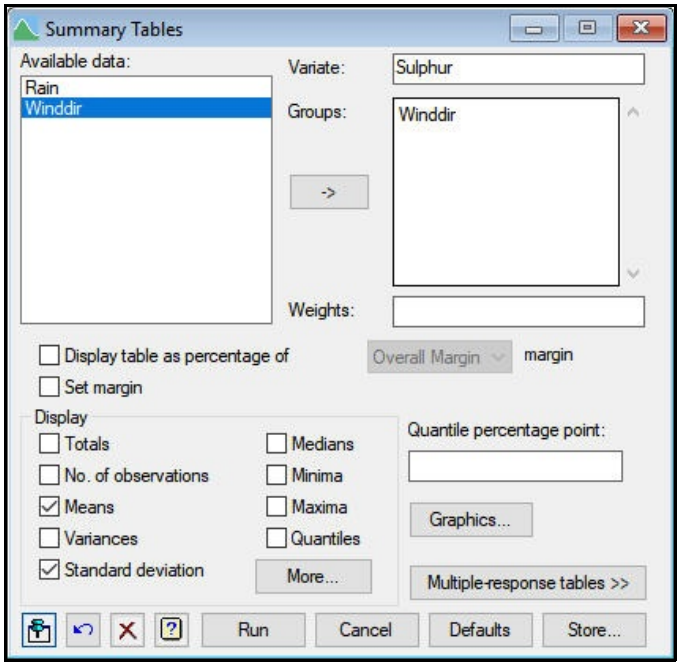


Figure 4.18

In Figure 4.18 we have selected the Means and Standard deviation boxes, selected `Sulphur` as the Variate and `Winddir` for the Groups. Clicking on Run produces the following table.

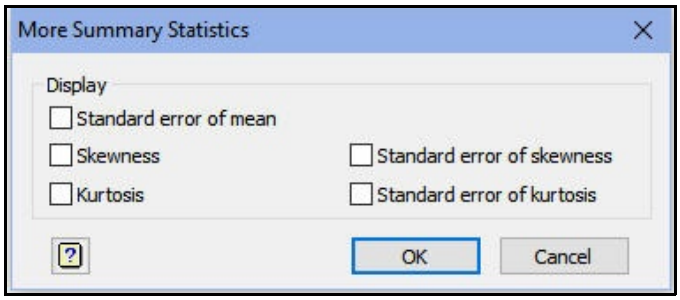


Figure 4.19

Mean		s.d.
Winddir		
E	9.88	5.866
N	18.88	15.236
NE	14.27	11.748
NW	17.27	12.059
S	8.13	9.591
SE	11.36	12.460
SW	4.71	3.053
W	10.81	9.532

Unknown cell

	Mean	10.00

This shows two one-way tables for the means and their (within-cell) standard deviations printed in parallel. As with the Frequency Tables menu, you can form up to nine-way summary tables. Tables of summaries can be saved using the Summary Tables Store Options menu, obtained by clicking the Store button.

4.5 Practical

File `Pet.gsh` contains a small school survey on children with pets. The three columns indicate the sex of the child (boy or girl), whether or not they have a pet, and their ages. Produce a two-way table showing the mean ages of the children of each sex with and without pets.

4.6 Association between categorical variables

Relationships between categorical variables cannot be displayed in the same way as between continuous variables. For example, it would not be possible to draw a useful scatter plot of wind direction against rain status. However, it is still reasonable to look for relationships between such variables.

One way to display the relationships is to tabulate the numbers of observations in the categories. In Figure 4.20 we show how to tabulate the number of rainy and dry days for each wind direction using the Frequency Tables menu.

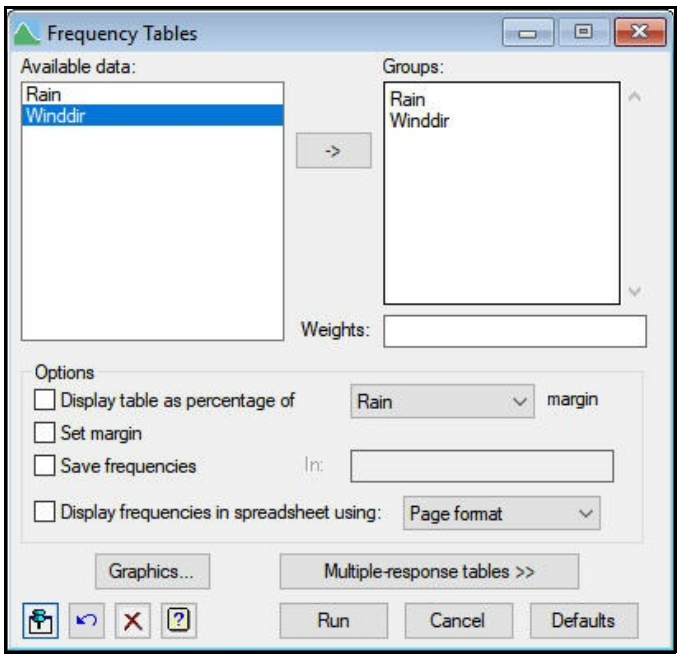


Figure 4.20

Count

Winddir Rain	E	N	NE	NW	S
no	7	5	5	8	5
yes	1	3	6	3	10

Winddir Rain	SE	SW	W
no	7	14	13
yes	4	14	8

```
Unknown cell
             Count                1
```

This indicates that there appears to be a higher proportion of rainy days when the wind is from a southerly direction. Notice that Genstat automatically divides the display into sections if it is too wide to be shown in a single array.

The layout of the display is also affected by the order of the factors listed in the Groups box. In this example, a clearer display is achieved by reversing the order (`Winddir` and then `Rain`): if there is more than one factor, Genstat displays categories of the last factor across the page, and all the others down the page.

```
Count
Rain          no        yes
Winddir
E              7          1
N              5          3
NE             5          6
NW             8          3
S              5         10
SE             7          4
SW            14         14
W             13          8

Unknown cell
             Count                1
```

If you wished to test the apparent association between these two factors, you could carry out a *chi-square test*. This attempts to evaluate whether there are any significant differences in the proportion of rainy days for each wind direction; or, equivalently, whether there is a significantly different distribution of directions on wet and dry days.

To produce a chi-square test you first need to click Stats on the menu bar, then click on Statistical Tests and select Contingency Tables (Chi-square). The type of test is selected using the drop-down list. When you select Chi-square test, the menu takes the form shown in Figure 4.21. The Data arrangement option allows you to specify the form of the data. You can provide the data in a two-way table; this can be saved using the Frequency Table menu or you can create a table by clicking on the Create table button. Alternatively, you can supply the row and column factors, or a single variate containing the counts with two grouping factors to identify the row and columns. The Method option

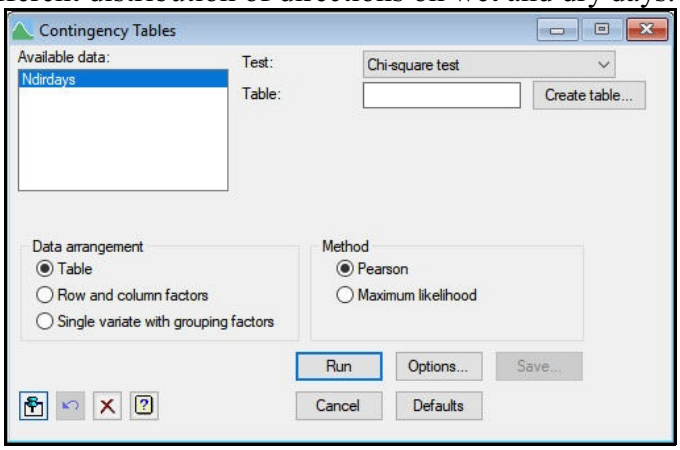


Figure 4.21

allows you to select between two-ways of calculating the test: Pearson uses the familiar method; the alternative, which may be more accurate, uses maximum likelihood.

In Figure 4.22 we have selected the Pearson method and the Row and column factors data arrangement. We have entered `Winddir` as the row factor, `Rain` as the column factor and the name `RainDirCount` to save the resulting two-way table. Clicking the Run button produces the output shown below.

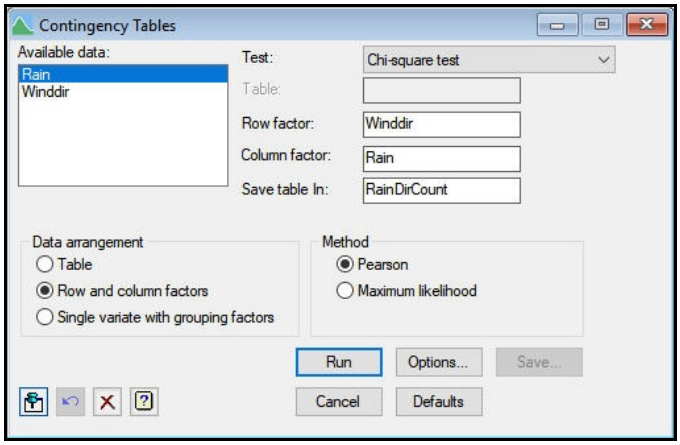


Figure 4.22

Chi-square test for association between Winddir and Rain

Message: some expected values less than 5, so test may be unreliable.

Pearson chi-square value is 9.21 with 7 d.f.

Probability level (under null hypothesis) p = 0.238

The output contains a message to warn that at least one of the factor combinations (or cells) has an expected value less than 5. This means that the chi-square test may not be valid, since it is based on an approximation that is valid only for large numbers. If it were valid, the probability is too large to conclude a significant association even at the 20% significance level. A general rule of thumb with a chi-square test is that if less than 20% of the values have expected values less than 5, and all have expected values greater than 1, then you can cautiously accept the results of the chi-square test.

You can view the expected values by selecting the Expected values display option from the Chi-Square Options menu (Figure 4.23), obtained as usual by clicking on the Options button in the Contingency Tables menu.

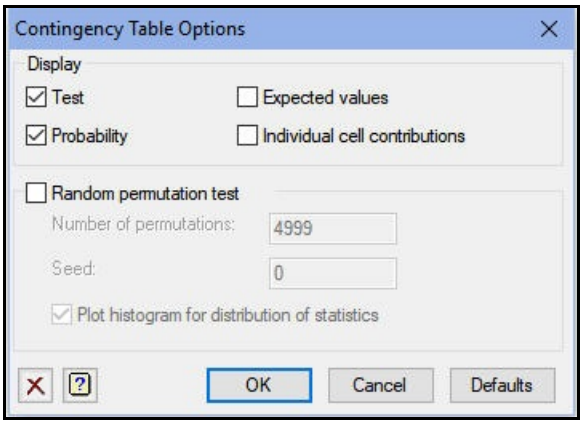


Figure 4.23

Alternatively, you can display the expected values in a spreadsheet. Click on the Save button in the Contingency Tables menu, to open the Contingency Table Save Options menu shown in Figure 4.24. Selecting the box for Expected values will enable a window (entitled In:) into which you enter the identifier of the structure in which the information is to be saved. In Figure 4.24 we have selected this box and entered the name ExpectedVals. To display this structure in a spreadsheet we have selected the Display in spreadsheet box. Clicking on Save forms the table ExpectedVals, and puts it into a spreadsheet as shown in Figure 4.25.

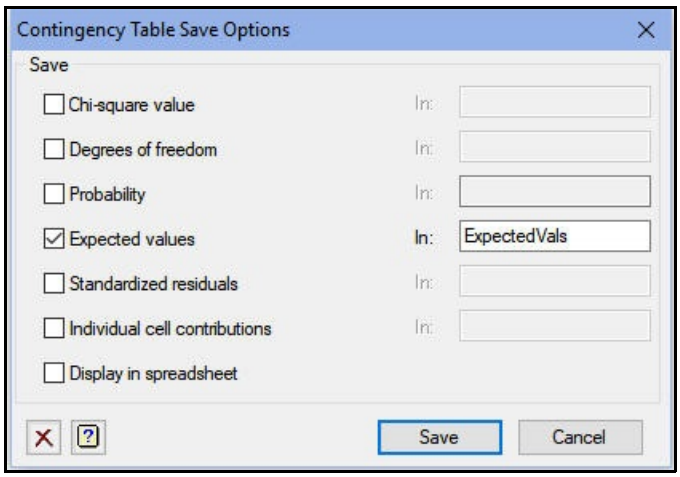


Figure 4.24

You can perform a chi-square test on a one-way table of counts using the Chi-square Goodness-of-fit menu (accessible through the Statistical Tests menu). In this case the test assesses whether the values in the various cells of the table are different; for example, we could test to see whether all the wind directions are roughly equally represented.

The Contingency Table menu also offers Fisher's exact test of association between the factors classifying a 2×2 table of counts. (This is the other option in the Test drop-down list box at the top of the menu.)

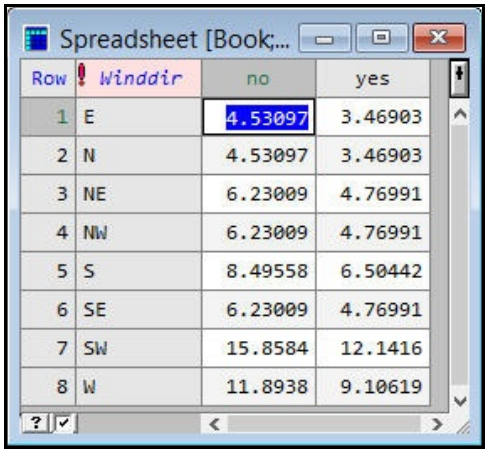


Figure 4.25

You can also investigate more general relationships between categorical variables using *log-linear models*, as described in Section 5.7.

4.7 Practical

Using the school survey on children with pets (Pet.gsh) from Practical 4.5, carry out a chi-square test of the hypothesis that equal proportions of boys and girls have no pets. Use the Create table button to form a new two-way table from the chi-square menu to save the numbers of girls and boys with and without pets.

4.8 Transferring output to other applications

When writing a report from an analysis, you may want to include some Genstat output within the text. One way of transferring this information is to use the standard Cut, Copy and Paste options available in the Edit menu. This is most effective when the Output window is displaying the information in rich-text style. As explained in Section 1.1,

columns of output are then separated from each other by tab characters. The formatting is thus preserved when you copy or paste the output into a word-processed document. As an alternative, we now show how you can form a table, in an application such as MS Word or MS Explorer, from information in a Genstat spreadsheet or text window.

To illustrate this we will copy the table of counts for wind direction and rain from the sulphur data, and insert it into a table within MS Word. We first need to load the table into a spreadsheet, which you can do by selecting the Data in Genstat sub-option of the New option of the Spread menu. This produces the menu shown in Figure 4.26, where we have selected the Table option from the Type of spreadsheet radio buttons, and double-clicked on the RainDirCount data structure to enter it into the Data to load field.

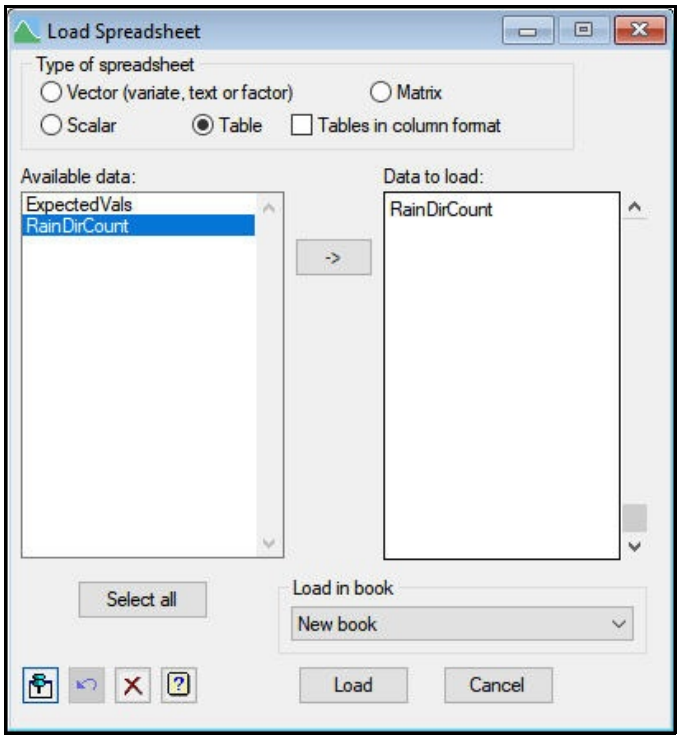
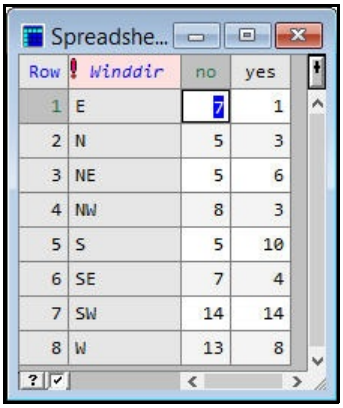


Figure 4.26

Clicking on Load produces the spreadsheet shown in Figure 4.27. So, the table that we wish to copy will have 3 columns and 9 rows, including the row and column titles. To produce a table in MS Word, you select the RTF Table option from the Copy Special option in the Edit menu on the menu bar.



Row	Winddir	no	yes
1	E	7	1
2	N	5	3
3	NE	5	6
4	NW	8	3
5	S	5	10
6	SE	7	4
7	SW	14	14
8	W	13	8

Figure 4.27

This produces the menu in Figure 4.28, which allows you to specify how the table is to appear in your Word document. In this example, we use the default settings for the formatting, and the variable names as column headings. Clicking OK now copies the table to the clipboard. In the Word document, you put the cursor in the position where you want the table to appear, and select Paste from the Edit menu on the menu bar. The table then appears in Word in the format shown in Figure 4.29, and you now can change the table attributes within Word as you like.

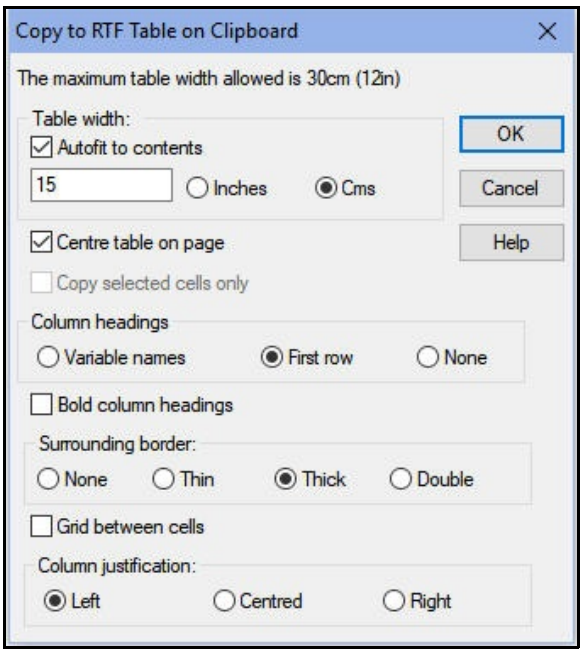


Figure 4.28

Winddir	no	yes
E	7	1
N	5	3
NE	5	6
NW	8	3
S	5	10
SE	7	4
SW	14	14
W	13	8

Figure 4.29

In this first example we have copied from a spreadsheet of type table, however, you can also apply this method to other types of spreadsheet, such as vector or scalar.

In our second example we copy a table from the Output window into MS Word. To display the table in the Output window we first open the Display Data in Output menu by clicking in the Display Data in Output option of the Data menu on the menu bar. We enter `RainDirCount` into the right-hand window, enter 12 as the field width and 0 as the number of decimal places, and then click on Apply (see Figure 4.30). Clicking on Run then produces the output shown in Figure 4.31.

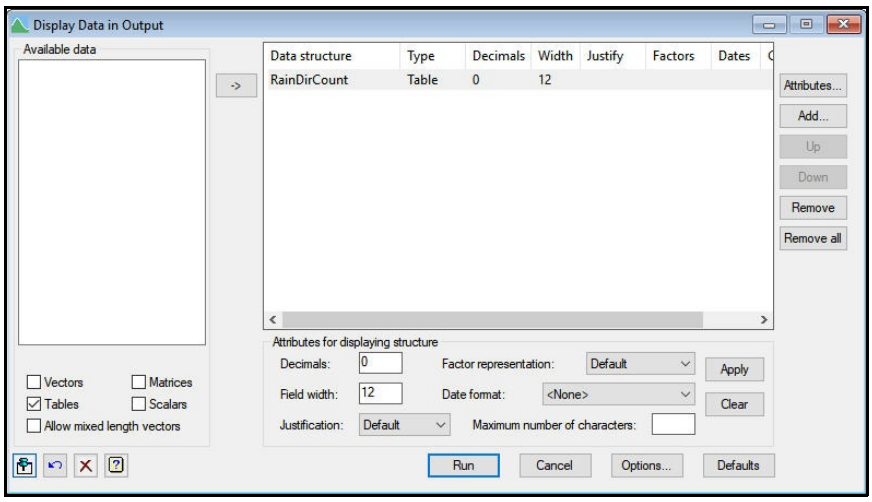


Figure 4.30

You can now select the output to appear in the table by using the mouse in the usual way. Alternatively, you can make a selection by holding the shift key down and pressing the arrow keys. Figure 4.31 shows the selected rows of the output that are to be transferred into the table.

If you have set the output viewer to display in RTF (see Sections 1.2 and 1.7), you can copy the output directly into Word with all the formatting preserved.

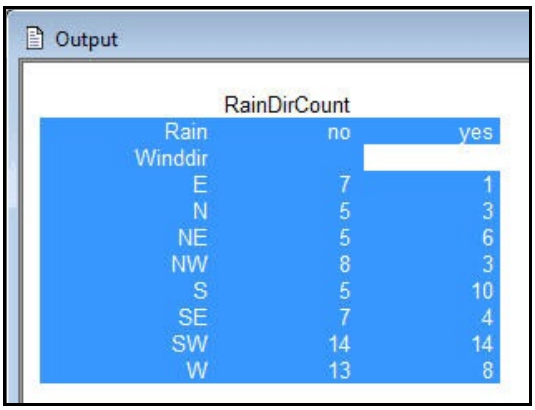


Figure 4.31

If you are displaying the output as plain text, you should copy the output into Word by selecting RTF Table from the Copy Special option on the Edit menu, which produces the menu shown in Figure 4.32. In the plain-text output style (as in Figure 4.31), spaces are used to separate different cells in a table. To produce RTF tables, the spaces between the columns must be converted to tab characters. In Figure 4.32 we have requested to convert columns with 2 or more spaces to tabs.

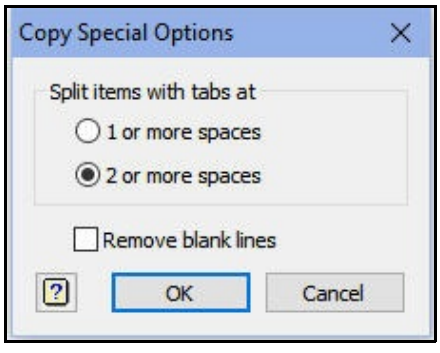


Figure 4.32

Clicking on OK then opens the menu to control the appearance of the table (Figure 4.33). Leaving the default settings and clicking on OK copies the table to the clipboard. In Word, you put the cursor in the position where the table is to appear, and select Paste from the Edit menu. Figure 4.34 shows the resulting table in Word.

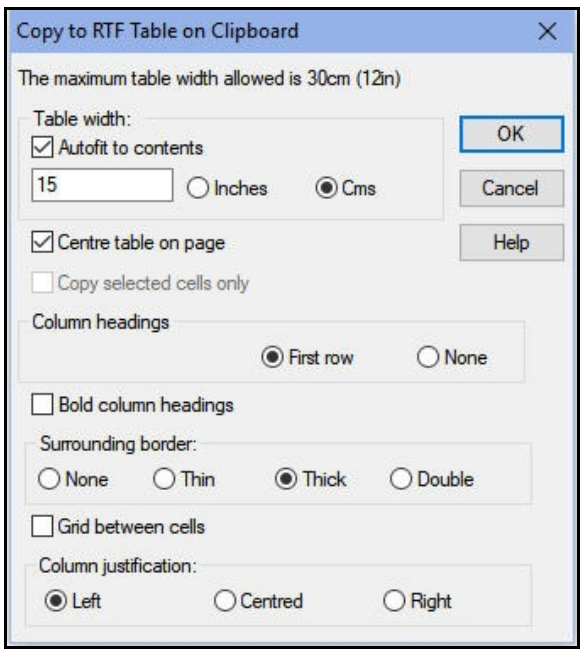


Figure 4.33

Rain	no	yes
Winddir		
E	7	1
N	5	3
NE	5	6
NW	8	3
S	5	10
SE	7	4
SW	14	14
W	13	8

Figure 4.34

A final possibility is that Genstat can rerun your analyses, and generate an output file in either RTF, HTML or LaTeX (or plain text). You first need to create a program file. If you want to rerun the entire session, make the Input Log the active window, by selecting the Input Log option in the Window menu on the menu bar. The save the log in a file, using the Save As option in the File menu on the menu bar. Alternatively, if you want to run just part of the session (and if you are confident using commands), cut and paste from the Input Log into a new text window, and save that.

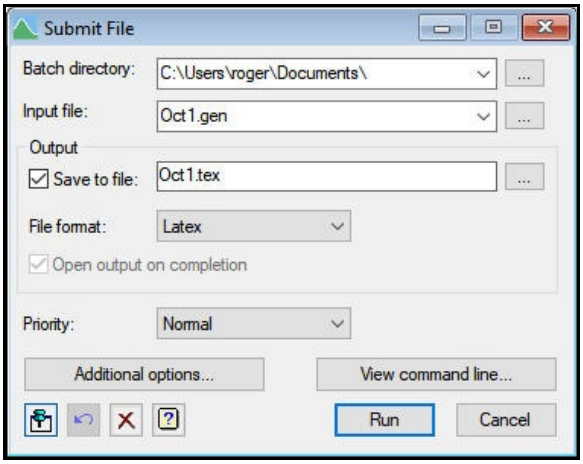


Figure 4.35

Now click on Run on the menu bar and select the Submit File option, to obtain the menu in Figure 4.35. The folder (or directory) where the program file is stored should be entered in the Batch directory box, the name of the program file should be entered into the Input file box, and the name of the output file into the Output file box. The button to the right of each box can be used to browse to find the folder or file. Here we have a program file called `Oct1.gen` in the folder `C:\Users\roger\Documents\Data` and are creating an output file `Oct1.tex` in LaTeX format.

4.9 Practical

In Practical 4.7 you formed a two-way table of counts of girls and boys with and without pets, formed from the school survey in `Pet.gsh`. Copy the table to the clipboard using the Copy Special options and paste the table into Word.

4.10 Commands for basic statistics

You can also generate the analyses and graphs described above directly using Genstat commands.

The bar chart in Figure 4.17 can be produced by the `BARCHART` directive as follows

```
BARCHART Ndirdays
```

The boxplots can be obtained using the `BOXPLOT` procedure. For example,

```
BOXPLOT [TITLE='sulphur pollution';\
        AXISTITLE='logarithm of sulphur']\
        LogSulphur; GROUPS=Rain
```

produces the picture in Figure 4.11. The `METHOD` option controls the type of boxplot, the `TITLE` option supplies a general title for the picture and the `AXISTITLE` option gives the title for the axis. The `GROUPS` parameter indicates that separate boxplots are to be produced for each level of the factor `Rain`. Boxplots can be also drawn in line-printer style by setting option `GRAPHICS=lineprinter`.

Descriptive statistics are produced by the `DESCRIBE` procedure. The default output is the same as that generated by the Summary of Variates in Figure 4.3. The `SELECTION` option controls what statistics are given; for example, you could specify

```
DESCRIBE [SELECTION=mean,var,skew,kurtosis] Sulphur
```

to get the variance, skewness and kurtosis of the sample as well as the mean, and exclude the other information shown by default. To produce a summary using groups you need to specify a factor in the `GROUPS` option; for example

```
DESCRIBE [SELECTION=mean,var,skew,kurtosis; \
         GROUPS=Rain] Sulphur
```

The frequency tables are generated by the `TABULATE` directive. The menu in Figure 4.15 generates the command

```
TABULATE [PRINT=count; CLASSIFICATION=Winddir; \
         COUNTS=Ndirdays]
```

and that in Figure 4.20 generates

```
TABULATE [PRINT=count; CLASSIFICATION=Rain,Winddir]
```

The `CLASSIFICATION` option specifies the factors that are to be used to classify the table, and the `PRINT` option indicates what output is to be printed. You can also produce tabular summaries of the values within a variate as shown in the Summary by Groups menu (Figure 4.18). For example,

```
TABULATE [PRINT=mean; CLASSIFICATION=Rain] Sulphur
```

`TABULATE` also has options and parameters that allow the information to be saved in Genstat table structures. The command

```
TABULATE [CLASSIFICATION=Rain,Winddir] Sulphur; \
```

```
      NOBSERVATIONS=RainDirCount
```

saves the numbers of observations in each of the categories of rain and wind direction in a table called `RainDirCount`. The `CHISQUARE` procedure can then be used to perform a chi-square test of association between the factors (Figure 4.22):

```
CHISQUARE RainDirCount
```

The Mann-Whitney test is performed by the `MANNWHITNEY` procedure. For example,

```
MANNWHITNEY [GROUPS=Rain] Sulphur
```

compares the values of sulphur in the two groups defined by the factor `rain` (see Figure 4.14). Similarly, t-tests are performed by the `TTEST` procedure:

```
TTEST [METHOD=twosided; CIPROB=0.95; VMETHOD=automatic; \
      GROUPS=Rain] LogSulphur
```

compares the values in the variate `LogSulphur` (Figure 4.12). With both of these procedures, as in the menus, you can also specify the data values for the groups in two separate variates, for example,

```
TTEST sample1; sample2
```

to compare the values in the variates `sample1` and `sample2`.

The values of sulphur are transformed using the `CALCULATE` directive and the standard logarithmic function (`LOG10`):

```
CALCULATE LogSulphur = LOG10(Sulphur)
```

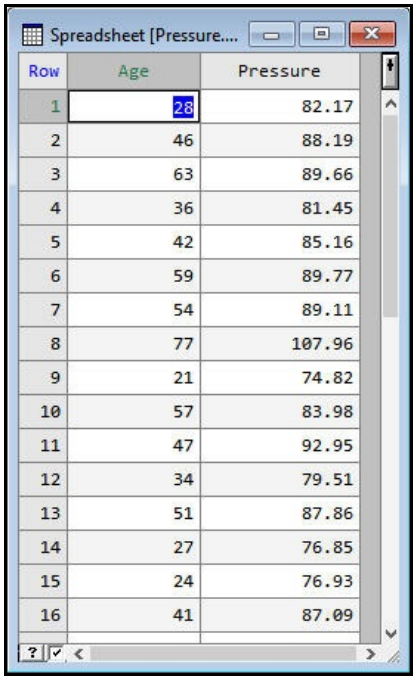
5 Regression

This chapter introduces the menus for fitting regression models in Genstat. We start with *simple linear regression*, where a straight line is fitted to represent the relationship between two variables; one variable is considered as the *response variable* (or *y-variate*) and the model predicts its mean value given the value of the other, *explanatory variable* (or *x-variate*). We then show how you can fit *parallel* and *non-parallel regressions* when you have an explanatory factor as well as an x-variate. We also introduce the Standard Curves menu, which can be used to fit a range of commonly-occurring nonlinear models. Finally we show how you can fit *generalized linear models*, when you have data like counts and proportions that cannot follow the usual assumption that the data come from Normal distributions.

A more comprehensive description of the regression menus is given in the *Guide to Regression, Nonlinear and Generalized Linear Models in Genstat*, while further details of the commands and the underlying statistical theory are in the *Guide to the Genstat Command Language, Part 2 Statistics*, Chapter 3. These can both be accessed from within Genstat *for Windows* by selecting sub-options of the Genstat Guides option of the Help menu on the menu bar; see Figure 1.6).

5.1 Simple linear regression

Spreadsheet file `Pressure.gsh` (Figure 5.1) contains recordings of the blood-pressure of a sample of 38 women whose ages range from 20 to 80. The file can be opened from within Genstat using the Example Data Sets menu, as explained in Section 4.1.



Row	Age	Pressure
1	28	82.17
2	46	88.19
3	63	89.66
4	36	81.45
5	42	85.16
6	59	89.77
7	54	89.11
8	77	107.96
9	21	74.82
10	57	83.98
11	47	92.95
12	34	79.51
13	51	87.86
14	27	76.85
15	24	76.93
16	41	87.09

Figure 5.1

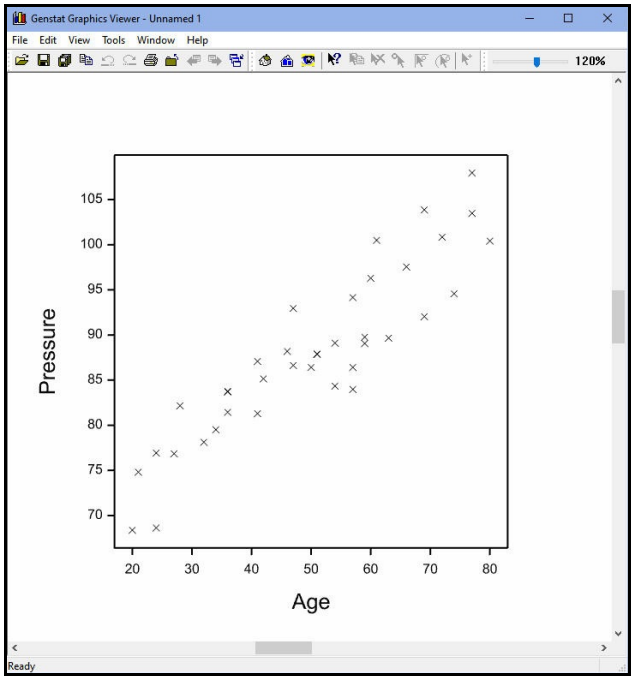


Figure 5.2

Figure 5.2 plots a graph of pressure against age (drawn by selecting 2D Scatter Plot option of the Graphics menu on the menu bar). This suggests that there is a linear relationship between blood-pressure and age. We will quantify this by a linear regression model, which specifies a *line of best fit* or a *regression line* between the points on the graph. It is natural here to assume that the blood-pressure is *responding* to increasing age, so we will fit a line or model to predict blood-pressure from age. The equation of the line is

$$pressure_i = a + b \times age_i + e_i$$

where a can be interpreted as the intercept of the regression line, b as its slope and e_i as the error, or vertical distance of the ith point from the line. A regression analysis produces estimates of the *parameters* a and b of this model, and also of the variance of the variable e which is often of as much interest as the parameters. Further information about method of estimation, and of the assumptions that are necessary, is given in Chapter 1 of the *Guide to Regression, Nonlinear Models and Generalized Linear Models*.

The model can be fitted in Genstat by selecting Regression Analysis from the Stats menu, and then clicking on Linear Models as shown in Figure 5.3.

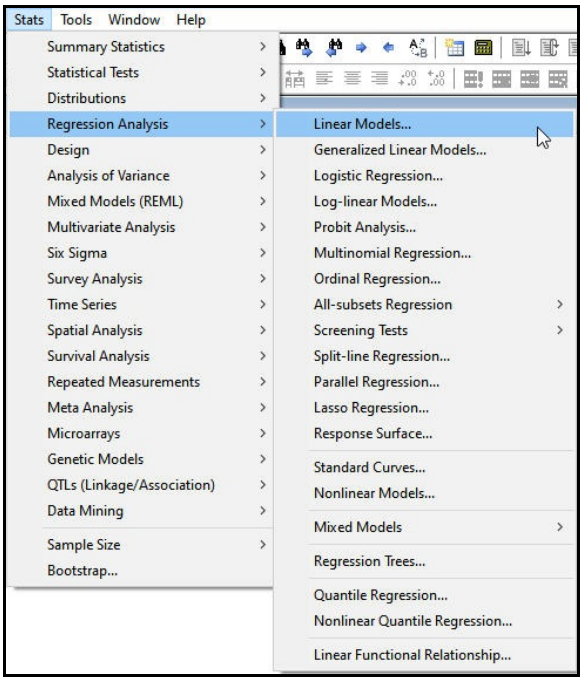


Figure 5.3

This opens the Linear Regression menu shown in Figure 5.4. If we select Simple Linear Regression in the drop-down list at the top, the menu customizes itself so that we just need to fill the Response variate and Explanatory variate boxes. We can then click on Run to produce the output below.

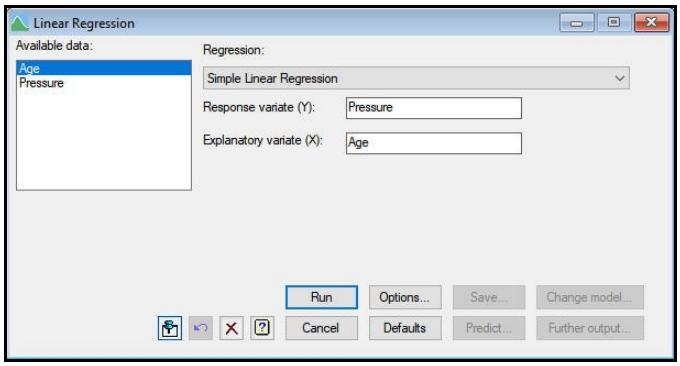


Figure 5.4

Regression analysis

Response variate: Pressure
Fitted terms: Constant, Age

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	2647.7	2647.69	169.73	<.001
Residual	36	561.6	15.60		
Total	37	3209.3	86.74		

Percentage variance accounted for 82.0
Standard error of observations is estimated to be 3.95.

Estimates of parameters

Parameter	estimate	s.e.	t(36)	t pr.
Constant	63.04	2.02	31.27	<.001
Age	0.4983	0.0382	13.03	<.001

The output begins with a description of the model, listing the response variable and the fitted terms: these are the explanatory variable and the *constant* or intercept term. The Linear Regression menu includes the constant by default; if you want to omit it, you can click on Options to open the Linear Regression Options menu (Figure 5.5), and uncheck the box Estimate constant term. This would constrain the fitted line to pass through the origin (that is, the response is zero when the explanatory is zero). However, this may be unwise if the data are close to the origin, as the analysis would still be based on the assumptions that the variability about the line is constant for the whole range of the data, and that the relationship is linear right down to the origin. The Options menu also allows you to select what output to display when the analysis is run. Figure 5.5 shows the default settings used to generate the output above.

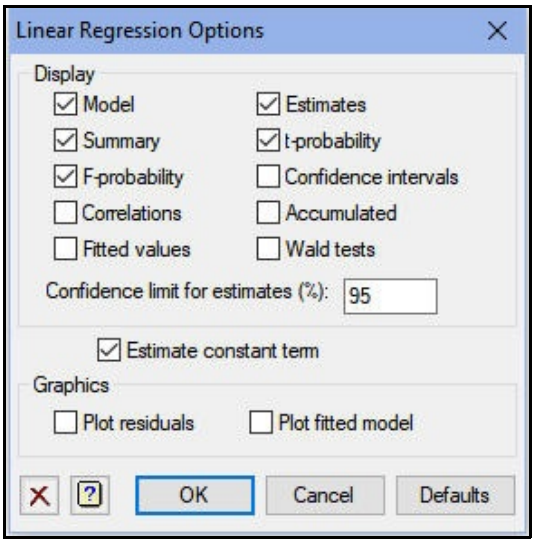


Figure 5.5

The Summary of Analysis contains an analysis of variance to assess the regression. The column headed *m.s.* (mean square) shows how much of the variance of the observations can be explained by the linear dependence on age (Regression), and how much is left over (Residual). The variance ratio (*v.r.*) is the ratio of the mean squares, and can be used to test formally whether there is a significant linear relationship. The column headed *F pr.* gives the probability of a variance ratio as large as this occurring by chance if there

were no relationship between the variables – but remember that this is based on the standard assumptions of linear regression, described later. A variance ratio as large as the one in this analysis indicates a significant relationship at the 0.1% level of significance (corresponding to probability 0.001).

The *percentage variance accounted for* is a summary of how much of the variability of this set of response measurements can be explained by the fitted model. It is the difference between residual and total mean squares expressed as a percentage of the total mean square. When expressed as a proportion rather than a percentage, this statistic is called the *adjusted* R^2; it is not quite the same as R^2, the squared coefficient of correlation. The adjustment takes account of the number of parameters in the model compared to the number of observations.

The final section of the analysis, displays the estimates of the parameters in the model. So, for example, you can see that blood-pressure rises on average by 0.4983 units for each year of age, with a standard error of 0.0382. The corresponding t-statistic is large, 13.03 with 36 degrees of freedom, indicating that there is a significant association between pressure and age as we expected from the graph. Again, the significance level is based on the standard assumptions of linear regression.

The Regression Options menu (Figure 5.5), allows you to ask for specific sections of output before the analysis is carried out. Alternatively, after the analysis, you can click on Further Output in the Linear Regression menu and ask for other sections. The resulting Linear Regression Further Output menu is shown in Figure 5.6; if you check Summary or Estimates, further options are available so that you can choose whether probabilities should appear with the analysis.

The Fitted values selection produces the display shown below.

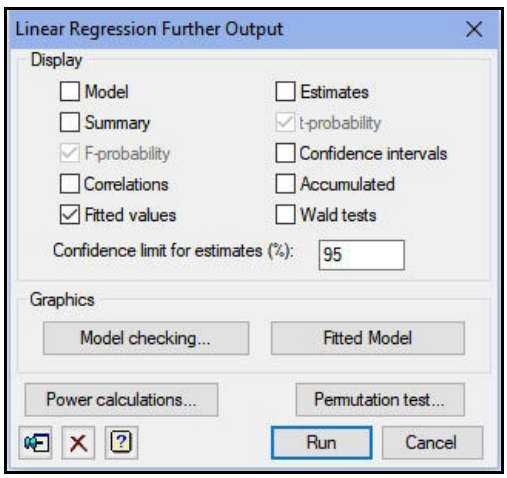


Figure 5.6

Regression analysis

Fitted values and residuals

Unit	Response	Fitted value	Standardized residual	Leverage
1	82.17	77.00	1.36	0.072
2	88.19	85.97	0.57	0.028
3	89.66	94.44	-1.24	0.042
4	81.45	80.98	0.12	0.045
5	85.16	83.97	0.31	0.032
6	89.77	92.44	-0.69	0.034
7	89.11	89.95	-0.22	0.028
8	107.96	101.41	1.74	0.095
9	74.82	73.51	0.35	0.105
10	83.98	91.45	-1.92	0.031

11	92.95	86.46	1.67	0.027
12	79.51	79.99	-0.12	0.050
13	87.86	88.46	-0.15	0.026
14	76.85	76.50	0.09	0.076
15	76.93	75.00	0.51	0.090
16	87.09	83.47	0.93	0.034
17	97.55	95.93	0.42	0.050
18	92.04	97.43	-1.41	0.060
19	100.85	98.92	0.51	0.072
20	96.30	92.94	0.87	0.036
21	86.42	87.96	-0.39	0.026
22	94.16	91.45	0.70	0.031
23	78.12	78.99	-0.23	0.057
24	89.06	92.44	-0.87	0.034
25	94.58	99.92	-1.41	0.080
26	103.48	101.41	0.55	0.095
27	81.30	83.47	-0.56	0.034
28	83.71	80.98	0.71	0.045
29	68.38	73.01	-1.24	0.111
30	86.64	86.46	0.05	0.027
31	87.91	88.46	-0.14	0.026
32	86.42	91.45	-1.29	0.031
33	103.87	97.43	1.68	0.060
34	83.76	80.98	0.72	0.045
35	84.35	89.95	-1.44	0.028
36	68.64	75.00	-1.69	0.090
37	100.50	93.44	1.82	0.038
38	100.42	102.91	-0.67	0.111
Mean	87.95	87.95	0.00	0.053

The *fitted values* are those predicted by the model for each observation; that is, $a + b \times x_i$. Instead of displaying the *simple residuals*, e_i, these values have been divided by their standard error: the resulting *standardized residuals* should be like observations from a Normal distribution with unit variance, if the assumptions made in this analysis are valid. The *leverage* values indicate how influential each observation is: a large value indicates that the fit of the model depends strongly on that observation.

You can display the fit graphically by clicking on the Fitted model button in the Linear Regression Further Output (Figure 5.6). This displays the picture shown in Figure 5.7, which shows the observed data and the fitted line together with 95% confidence limits for the line.

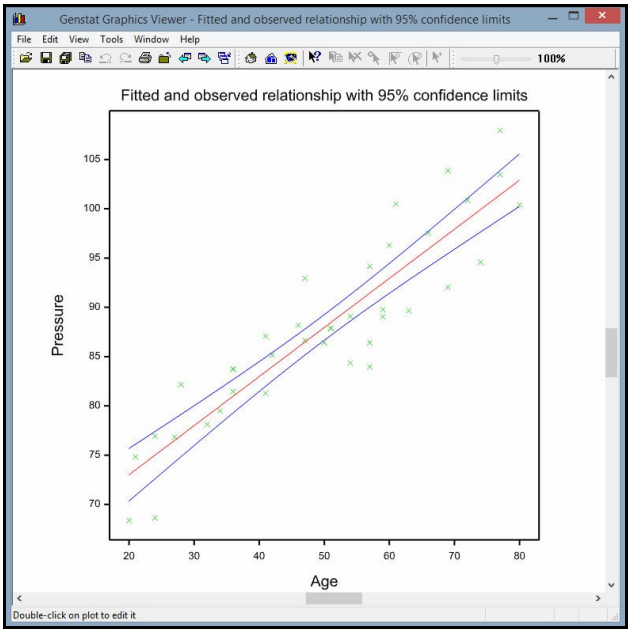


Figure 5.7

The Model Checking menu (Figure 5.8) allows you to plot various graphs to check the assumptions of the analysis visually. This can be obtained by clicking on the Model checking button in the Regression Further Output menu (Figure 5.6). The menu allows you to choose between five types of graph for either the residuals, the leverage values or the *Cook's statistics* (a combination of the residual and leverage information).

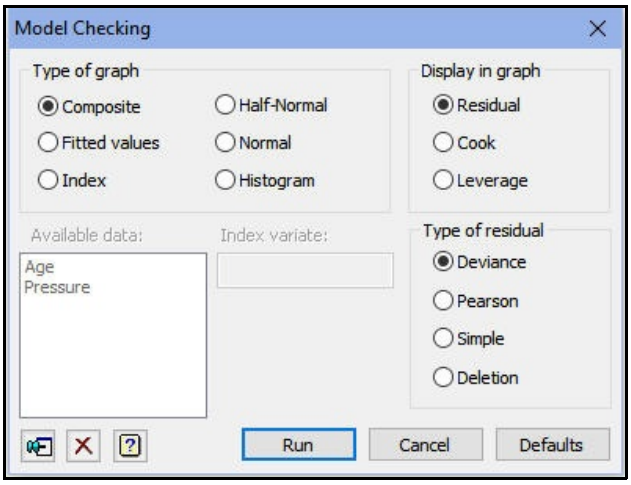


Figure 5.8

Figure 5.9 shows the default, which is a composite of four of these graphs: a histogram of the residuals, so that you can check that the distribution is symmetrical and reasonably Normal; a plot of residuals against fitted values, so that you can check whether the residuals are roughly symmetrically distributed with constant variance; a *Normal plot* which plots the ordered residuals against Normal distribution statistics – if they lie roughly on a straight line, the residuals are roughly Normally distributed; and a *half-Normal plot* which does the same for the absolute values of the residuals, and can be more useful for small sets of data.

These plots indicate that the variance seems unrelated to the size of the observation, but that the distribution seems to be more constrained than the Normal: the largest residuals are a little smaller than would be expected from a Normal distribution. Experience shows the analysis is robust to small departures from Normality. However, we should be cautious in interpreting the F-statistics and t-statistics (which rely on the assumption of Normality), if the histogram looks very non-Normal.

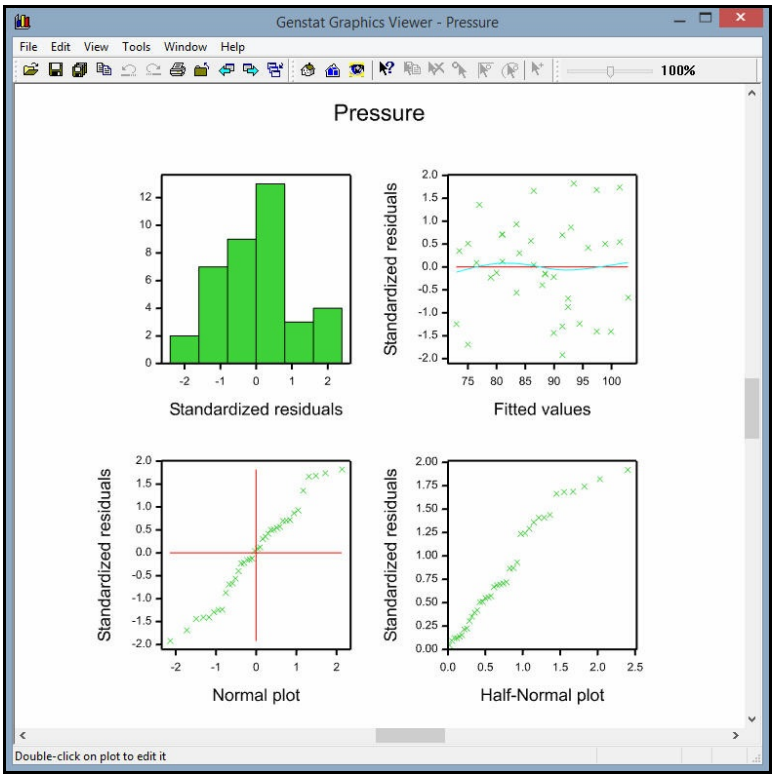


Figure 5.9

An alternative way to assess the significance of the regression might then be to use a permutation test. Clicking on the Permutation test button in the Linear Regression Further Output menu (Figure 5.6) produces the menu in Figure 5.10. This is asking Genstat to make 4999 random permutations of the values of the response variate (see the Number of permutations box), and refit the regression. The value 0 in the Seed box means automatically for the random numbers that are used to generate the permutations. If you have used random numbers already in the current run of Genstat, the seed will be chosen to continue the existing sequence. Otherwise it is initialized automatically (and the value is printed in the output). The probability for the regression is now determined from its distribution over the randomly permuted data sets. The output below shows a probability <.001, which means that the observed data set had one of the 5 largest variance ratios out of the 5000 sets that have been examined (1 observed data set + 4999 randomly permuted data sets).

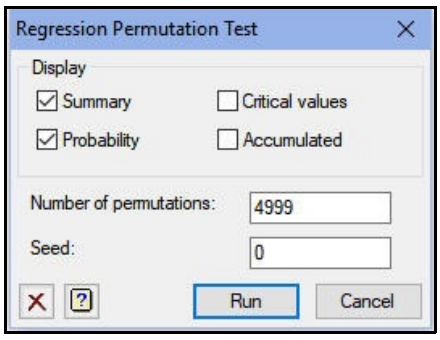


Figure 5.10

Message: Default seed for random number generator used with value 525069

Probability for model <.001 (determined from 4999 random permutations)

If you ask for more permutations than the number that are possible for your data, Genstat will instead do an *exact test*, which uses each permutation once. There are $n!$ (n factorial) permutations for a data set with n observations. So, we would obtain an exact test with 5 observations by setting the number of permutations to 120 or more.

As well as displaying the results of an analysis, the regression menus allow you to save the results in standard data structures. This is a common feature of most of the analysis menus in Genstat. After a regression analysis you can click on the Save button of the Linear Regression menu (Figure 5.4), to open the Linear Regression Save Options menu. The residuals, fitted values, parameter estimates and standard errors can all be saved in variates: if you check one of these boxes, you will be prompted for the name of the variate to store the results, as shown in Figure 5.11. The variance-covariance matrix of the parameter estimates can be saved in a symmetric matrix (another of Genstat's standard data structures).

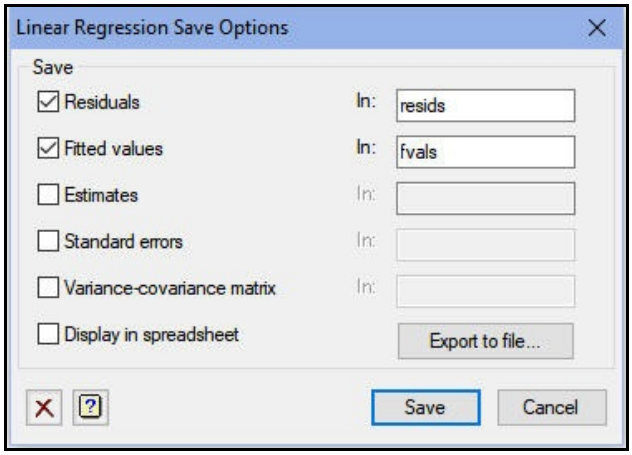


Figure 5.11

The fitted values provide predictions of the response variable at the values of the explanatory variable that actually occurred in the data. If you want predictions at other values, you can use the prediction menu, obtained by clicking on the Predict button in the Linear Regression menu. This generates the Predictions - Simple Linear Regression menu shown in Figure 5.12. Initially the Predict values at box has mean filled in, so that a prediction would be formed for pressure at the mean value of the ages. However, we have changed this to ask for predictions at ages 25, 50, 75 and 100. The Display box has boxes that can be checked to provide predictions, standard errors, standard errors of differences between predictions, least significant differences of predictions, confidence limits and a description of how the predictions are formed. Here we print predictions, standard errors and the description.

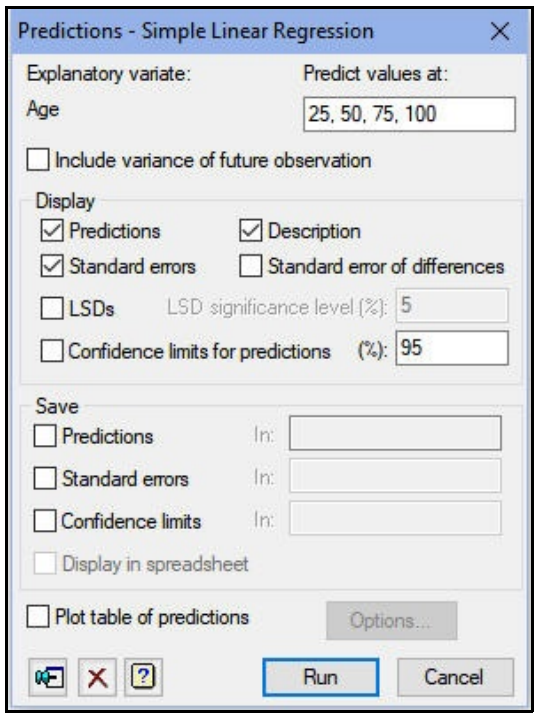


Figure 5.12

Predictions from regression model

These predictions are estimated mean values.

The standard errors are appropriate for interpretation of the predictions as summaries of the data rather than as forecasts of new observations.

Response variate: Pressure

Age	Prediction	s.e.
25	75.50	1.150
50	87.96	0.641
75	100.42	1.152
100	112.87	2.018

The output explains that the standard errors are appropriate as predictions for fitted values for these ages in this data set, not as predictions for new observations. We can augment the standard errors by the additional variability arising from a new set of observations at ages 25 - 100 by checking the box Include variance of future observation. (For further details see Section 3.3.4 of the *Guide to the Genstat Command Language, Part 2 Statistics*.)

5.2 Practical

An absorptiometer was used to measure the absorption of light passing through suspensions that contained different numbers of cells. It was intended to estimate the number of cells in future suspensions by the rapid light absorption method, so it was decided to fit a regression of cell counts on light absorption. The data are available in the spreadsheet file `Absorb.gsh`, where X is the absorptiometer reading and Y the cell count (10^8/ml). This example comes from *Experimentation in Biology* by Ridgman (1975, Blackie, Glasgow).

Load these data into Genstat and fit a linear regression of cell count on absorptiometer reading. Produce a graphical display of the regression. Compare this with a model with no constant (or intercept term).

5.3 Regression with groups

This section introduces the types of model that you can fit when you have factors in a regression model. Suppose you have one explanatory factor and one explanatory variate. You may then want to see how the regression line for the explanatory variate is the same within all the groups defined by the factor. Or perhaps the slope is the same for all the groups but the intercepts differ. Or perhaps the lines have different slopes and different intercepts.

We shall illustrate the possibilities using the sulphur pollution data in spreadsheet file `Sulphur.gsh`. from Chapter 3. First, we fit a simple linear regression on the wind speed.

Regression analysis

Response variate: Sulphur
Fitted terms: Constant, Windsp

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	935.	934.52	9.49	0.003
Residual	111	10932.	98.48		
Total	112	11866.	105.95		

Percentage variance accounted for 7.0
Standard error of observations is estimated to be 9.92.

Message: the following units have large standardized residuals.

Unit	Response	Residual
20	49.00	3.57
98	43.00	3.88

Message: the following units have high leverage.

Unit	Response	Leverage
30	3.00	0.075
72	5.00	0.051
95	14.00	0.054
100	25.00	0.051

Estimates of parameters

Parameter	estimate	s.e.	t(111)	t pr.
Constant	17.03	2.33	7.32	<.001
Windsp	-0.636	0.207	-3.08	0.003

We discovered in Section 3.1 that the sulphur measurements have a skewed distribution, so it is no surprise to find that the model checking plots in Figure 5.13 show a very skewed distribution of residuals. So, as in Section 4.1, we shall transform the sulphur measurements to logarithms, forming a new variate `LogSulphur` as before, containing the logarithms (to base 10) of the sulphur values.

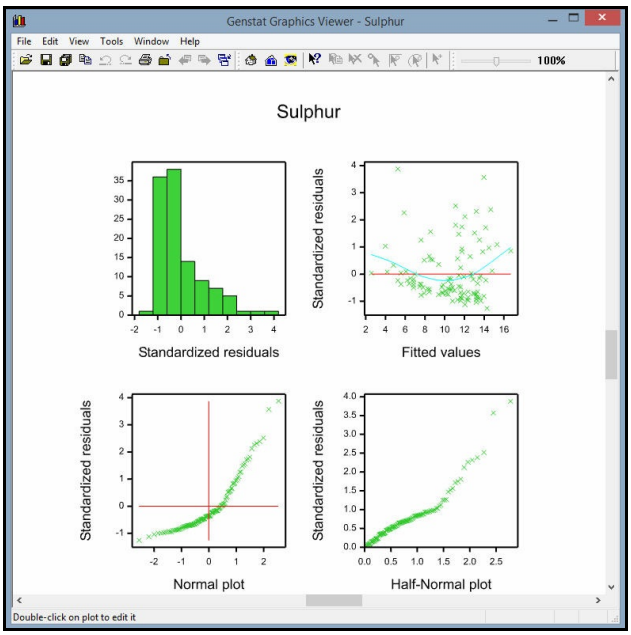


Figure 5.13

Regression analysis

Response variate: LogSulphur
Fitted terms: Constant, Windsp

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	1.50	1.4952	10.35	0.002
Residual	110	15.89	0.1445		
Total	111	17.39	0.1567		

Percentage variance accounted for 7.8
Standard error of observations is estimated to be 0.380.

Message: the following units have large standardized residuals.

Unit	Response	Residual
98	1.633	2.68

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.076
72	0.699	0.052
95	1.146	0.055
100	1.398	0.051

Estimates of parameters

Parameter	estimate	s.e.	t(110)	t pr.
Constant	1.1066	0.0892	12.41	<.001
Windsp	-0.02557	0.00795	-3.22	0.002

The residual plot in Figure 5.14 shows a much more symmetric distribution of residuals, with no evidence that the variance is changing with the size of the sulphur measurement. The plot does show up the imprecise recording of the sulphur measurements as integers: the apparent diagonal lines of points correspond to sulphur measurements with equal values.

The decrease in sulphur measurements with wind speed is estimated to be about 5.7% per km/h (antilog(−0.02557) = 94.3%), and is statistically significant.

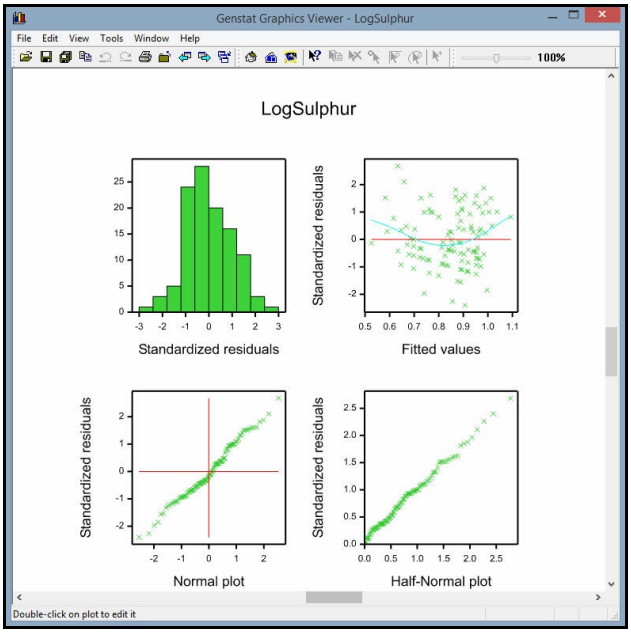


Figure 5.14

To investigate how the relationship is affected by rainfall, we select Simple linear regression with groups in the drop-down list at the top of the Linear Regression menu, as shown in Figure 5.15. This customizes a menu to include an extra box where we can specify a factor (here the factor `Rain`) to define groups to be investigated.

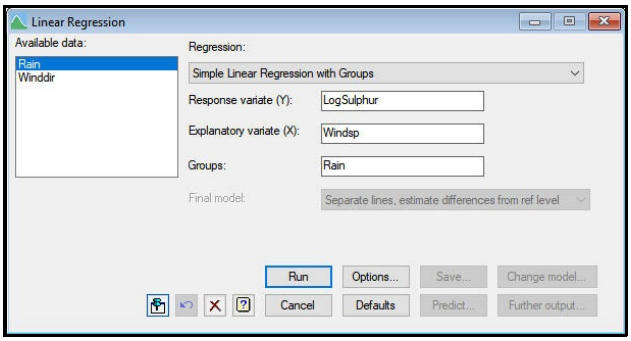


Figure 5.15

When we click on Run, three successive analyses are done. The first is exactly the same as that produced already with the Simple linear regression option, so we did not need to do that analysis separately. The second analysis fits a model with a separate intercept for wet and dry days, as shown below.

Regression analysis

Response variate: LogSulphur
Fitted terms: Constant + Windsp + Rain

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	1.89	0.9442	6.64	0.002
Residual	109	15.50	0.1422		
Total	111	17.39	0.1567		
Change	-1	-0.39	0.3933	2.77	0.099

Percentage variance accounted for 9.2
Standard error of observations is estimated to be 0.377.

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.102
72	0.699	0.073

Estimates of parameters

Parameter	estimate	s.e.	t(109)	t pr.
Constant	1.1235	0.0891	12.62	<.001
Windsp	-0.02193	0.00818	-2.68	0.008
Rain yes	-0.1240	0.0745	-1.66	0.099

Parameters for factors are differences compared with the reference level:

Factor	Reference level
Rain	no

The effect of rainfall is represented here by the difference between dry and wet days: that is, by comparing level `yes` of the factor `Rain` to its *reference level* `no`. (By default the reference level is the first level of the factor, but the Column Attributes spreadsheet menu allows you to choose other levels.) So the model is

$$LogSulphur = a + b \times Windsp$$

for dry days, and

$$LogSulphur = a + d + b \times Windsp$$

for wet days. The model thus consists of two *parallel* regression lines. The estimates show that rainfall decreases the sulphur on average by 25% (antilog(–0.1240) = 75%), but this effect is not statistically significant there is still a large amount of unexplained variation in the sulphur measurements. This version of the model is very convenient if you want to make comparisons with the reference level (which may, for example, represent a standard set of conditions or treatment). However, we show later in this section how you can obtain the alternative version with a parameter in the model for each intercept.

We can see whether the linear effect of wind speed is different in the two categories of rainfall by looking at the third and final analysis in the Output window.

Regression analysis

Response variate: LogSulphur
Fitted terms: Constant + Windsp + Rain + Windsp.Rain

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	3	1.92	0.6402	4.47	0.005
Residual	108	15.47	0.1432		
Total	111	17.39	0.1567		
Change	-1	-0.03	0.0323	0.23	0.636

Percentage variance accounted for 8.6
Standard error of observations is estimated to be 0.378.

Message: the following units have large standardized residuals.

Unit	Response	Residual
98	1.633	2.61

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.160
72	0.699	0.112
95	1.146	0.111
104	1.580	0.093

Estimates of parameters

Parameter	estimate	s.e.	t(108)	t pr.
Constant	1.153	0.109	10.57	<.001
Windsp	-0.0252	0.0107	-2.36	0.020
Rain yes	-0.208	0.193	-1.08	0.283
Windsp.Rain yes	0.0079	0.0167	0.47	0.636

Parameters for factors are differences compared with the reference level:

Factor	Reference level
Rain	no

This model includes the *interaction* between the explanatory factor and variate. In Genstat, interactions are represented using the dot operator, so that `Windsp.Rain` represents the interaction between wind speed and rain. More details about the model formulae that are used to specify statistical models in Genstat are given in Section 6.7.

The output now shows the slope of the regression for dry days, titled `Windsp`, and the difference in slopes between wet and dry, titled `Windsp.Rain yes`. So again we can see immediately that the difference between the slopes is small and not significant. The graph of the fitted model is shown in Figure 5.16.

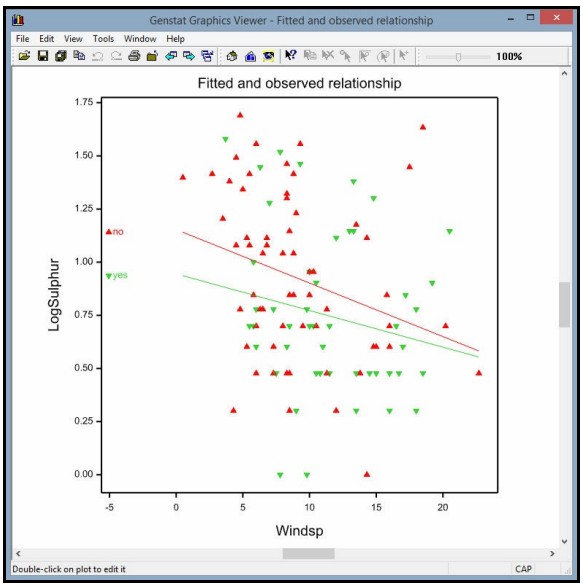


Figure 5.16

An *analysis of parallelism* can be carried out using the Accumulated option of the Linear Regression Further Output menu, as shown in Figure 5.17. This allows you to make a formal assessment of how complicated a model you need. You can then select the appropriate model from the Final model box in the Linear Regression menu (see Figure 5.18) and click on Run to fit it.

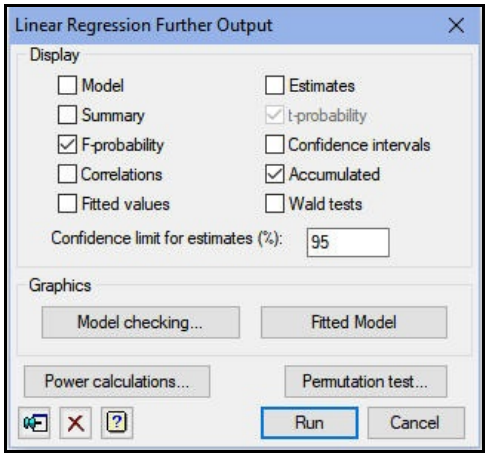


Figure 5.17

Regression analysis

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ Windsp	1	1.4952	1.4952	10.44	0.002
+ Rain	1	0.3933	0.3933	2.75	0.100
+ Windsp.Rain	1	0.0323	0.0323	0.23	0.636
Residual	108	15.4677	0.1432		
Total	111	17.3884	0.1567		

Here a Common line (in fact, a simple linear regression) would be enough, but to illustrate the fitted parallel lines we have selected Parallel lines, estimate lines in Figure 5.18.

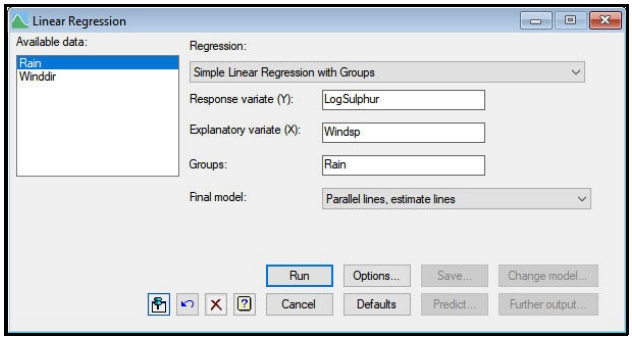


Figure 5.18

This fits parallel lines but now with a parameter for each intercept, rather than parameters for differences from the reference level (which would be given by the alternative setting Parallel lines, estimate differences from ref. level). The other settings are: Common line; Parallel lines, estimate differences from ref. level; Separate lines, estimate lines; and Separate lines, estimate differences from ref. level. The fitted parallel lines are shown in Figure 5.19.

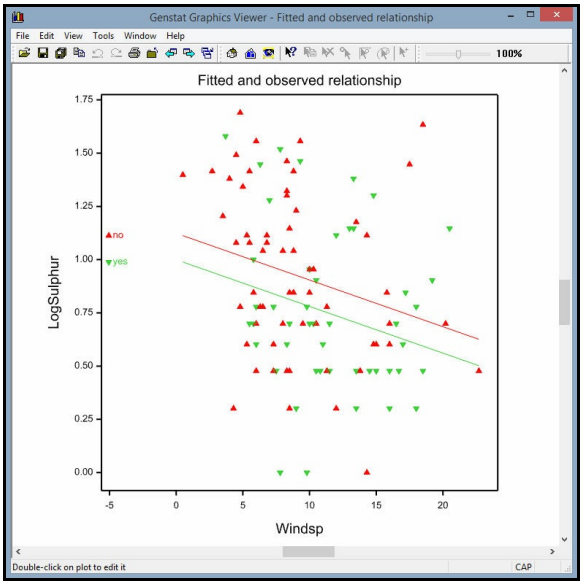


Figure 5.19

Regression analysis

Response variate: LogSulphur
Fitted terms: Windsp + Rain

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	1.89	0.9442	6.64	0.002
Residual	109	15.50	0.1422		
Total	111	17.39	0.1567		
Change	-1	-0.39	0.3933	2.77	0.099

Percentage variance accounted for 9.2
Standard error of observations is estimated to be 0.377.

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.102
72	0.699	0.073

Estimates of parameters

Parameter	estimate	s.e.	t(109)	t pr.
Windsp	-0.02193	0.00818	-2.68	0.008
Rain no	1.1235	0.0891	12.62	<.001
Rain yes	1.000	0.109	9.14	<.001

If we now click on the Predict button in the Linear Regression menu (Figure 5.22), we can obtain predictions from this parallel-line model. The predictions menu (Figure 5.20) is now customized to include the grouping factor (`Rain`).

In Figure 5.20, the drop-down list box Predict at levels is set to all, to indicate that we want to form predictions for all the levels of `Rain`. The alternative setting, standardize, forms averages over the levels of `Rain`, and the Standardization method box then allows you to indicate whether you want ordinary averages (Equal), or whether you want the levels weighted according to their replication in the data set (Marginal), or whether you want to specify your own weights (Specify) which might correspond to the numbers of wet and dry days that you would anticipate in some future period.

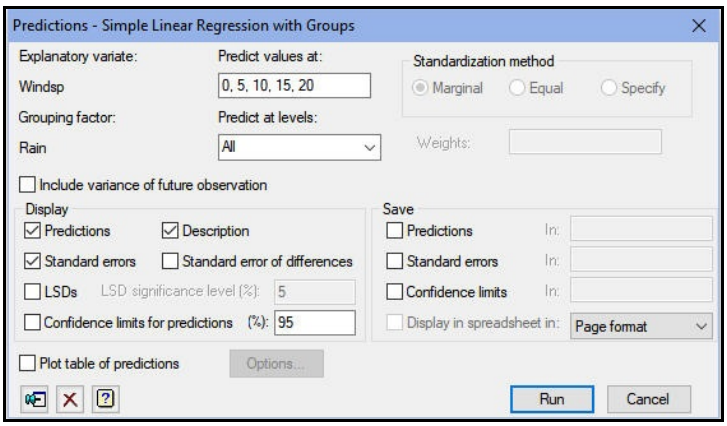


Figure 5.20

The other box specifies the values of the explanatory variate (`Windsp`) for which we want predictions, here 0, 5, 10, 15 and 20. We have also checked the box to include variance of future observation (unlike Figure 5.12 in Section 5.1), so the standard errors in the output below are relevant for the values as predictions of the amounts of sulphur on future occasions.

Predictions from regression model

These predictions are estimated mean values.

The predictions have been formed only for those combinations of factor levels for which means can be estimated without involving aliased parameters.

The standard errors are appropriate for interpretation of the predictions as forecasts of new observations rather than as summaries of the data.

Response variate: LogSulphur

Rain	no		yes	
	Prediction	s.e.	Prediction	s.e.
Windsp				
0	1.1235	0.3875	0.9996	0.3926
5	1.0138	0.3816	0.8899	0.3848
10	0.9042	0.3801	0.7802	0.3812
15	0.7945	0.3830	0.6705	0.3819
20	0.6848	0.3902	0.5609	0.3870

5.4 Practical

Experiments on cauliflowers in 1957 and 1958 provided data on the mean number of florets in the plant and the temperature during the growing season (expressed as accumulated temperature above 0°C).

1957		1958	
number	temp	number	temp
3.8	2.5	6.0	2.5
6.2	4.2	8.5	4.4
7.2	5.3	9.1	5.3
8.7	5.8	12.0	6.4
10.2	7.2	12.6	7.2
13.5	8.9	13.3	7.8
15.0	10.0	15.2	9.2

Load the data from file `Cauliflower.gsh` and carry out an analysis of parallelism of the relationship between number of florets and accumulated temperature, checking the assumptions for linear regression.

5.5 Fitting curves

In this section we show how to use the Standard Curves menu, which allows you to fit a range of commonly-occurring nonlinear models. We shall illustrate this using an experiment that was set up to study the relationship between yields of sugar cane and amounts of a nitrogen fertilizer. The data, in spreadsheet file `Cane.gsh`, consist of yields of sugar from four replicates of each of five amounts of the fertilizer.

A plot of the data, in Figure 5.21, shows a curved relationship, which we will model using an *exponential curve* (or *asymptotic regression*):

$$yield = a + b \times r^{nitrogen}$$

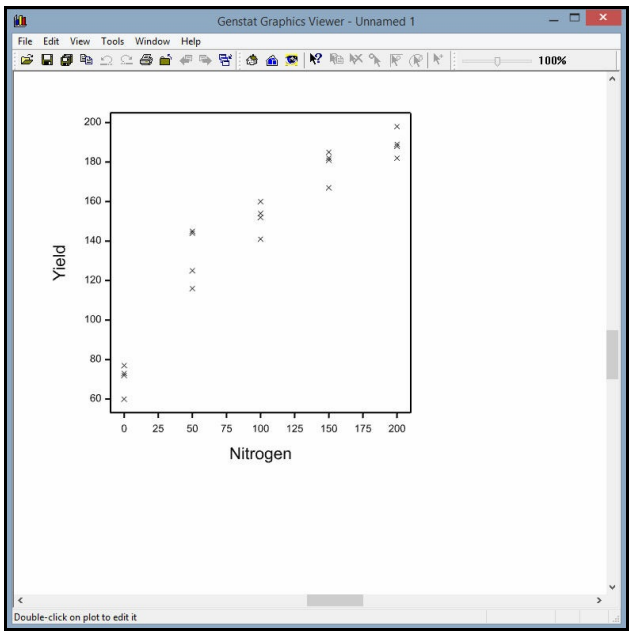


Figure 5.21

To open the Standard Curves menu (Figure 5.22) you need to select the Standard Curves sub-option of the Regression option of the Stats menu on the menu bar. The curve is selected from the list in the Type of curve box. To help you choose, the Example of Curve box changes to show a representative picture as you select each type. For the exponential with the Direction of response radio button set to Right, this shows a curve that rises to a plateau or *asymptote*, defined by the parameter *a*. The rate parameter *r* will be between 0 and 1, and the range parameter *b* will be negative.

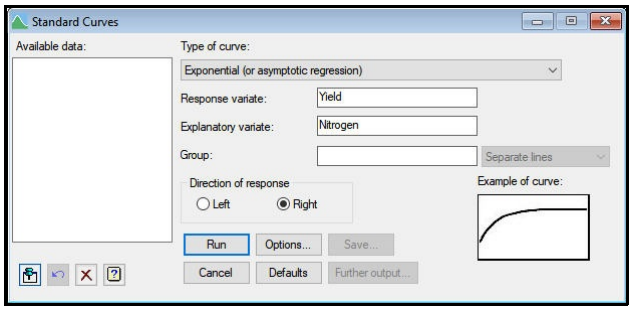


Figure 5.22

Here are the results of fitting the exponential curve to the sugar yields.

Nonlinear regression analysis

Response variate:	Yield
Explanatory:	Nitrogen
Fitted Curve:	A + B*(R**X)
Constraints:	R < 1

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	35046.	17523.18	182.02	<.001
Residual	17	1637.	96.27		
Total	19	36683.	1930.68		

Percentage variance accounted for 95.0
Standard error of observations is estimated to be 9.81.

Estimates of parameters

Parameter	estimate	s.e.
R	0.98920	0.00213
B	-131.1	10.6
A	203.0	10.8

Note that no t-probabilities are shown in this nonlinear analysis, because both the standard errors and the t-statistics are approximations, which depend on the amount of curvature of the model and on how well it fits the data.

The fitted model is shown in Figure 5.23. It seems to fit the data well, and has reasonable behaviour at both extremes of the nitrogen fertilizer treatments.

The Standard Curves menu covers most situations but, if you want to fit a curve that it does not cover, Genstat has an alternative menu, obtained by selecting the Nonlinear Models sub-option of the Regression option of the Stats menu, that allows you to define and fit your own nonlinear curves. Further details are given in Chapter 2 of the *Guide to Regression, Nonlinear and Generalized Linear Models in Genstat*.

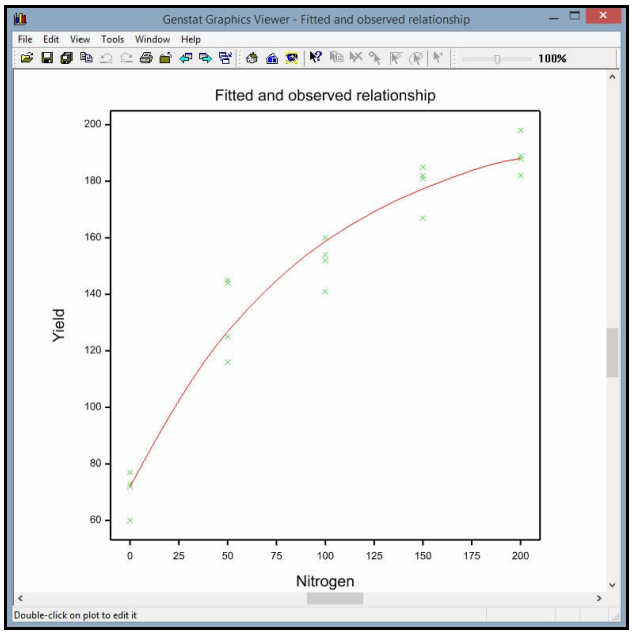


Figure 5.23

5.6 Practical

A product is known to lose weight after manufacture. The following measurements (in 1/16 oz) were taken at half-hourly intervals, and are available in file `Wtloss.GSH`:

Time after production	Weight difference
0.0	0.21
0.5	-1.46
1.0	-3.04
1.5	-3.21
2.0	-5.04
2.5	-5.37
3.0	-6.03
3.5	-7.21
4.0	-7.46
4.5	-7.96

This example comes from *Applied Regression Analysis* by Draper & Smith (1981, Wiley, New York).

Fit an exponential model to describe the loss of weight over time.

5.7 Generalized linear models

The regression menus that we have seen so far are intended for continuous data that can be assumed to follow a Normal distribution. However, Genstat can handle many other types of data.

One possibility is that the data may consist of counts. For example, you may have recorded the number of various types of items that have been sold in a shop, or numbers of accidents occurring on different types of road, or the number of fungal spores on plants with different spray treatments. Such data are generally assumed to follow a Poisson distribution. At the same time, it is usually assumed also that treatment effects will be proportionate (that is, the effect of a treatment will be to multiply the expected count by some number, rather than to increase it by some fixed amount). So, the model will be linear on a logarithmic scale rather than on the natural scale as used in ordinary linear regression. Models like this are known as *log-linear models* and form just one of the types of model covered by Genstat's facilities for generalized linear models.

We shall illustrate this with a data set (in Genstat spreadsheet file `Cans.gsh`) showing the number of cans of drink (`sales`) sold by a vending machine during 30 weeks. The explanatory variate `temperature` is the average temperature during that week.

The Generalized Linear Models menu is obtained by clicking on Generalized Linear sub-option of the Regression option of the Stats menu (see Figure 5.3). For a log-linear model, you should then select Log-linear modelling in the Analysis drop-down list box, as shown in Figure 5.24.

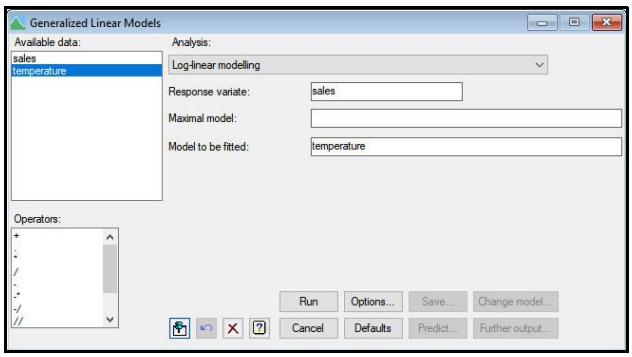


Figure 5.24

The menu has a box where you can specify a *maximal model*, as well as one where you can specify the model that you would like to fit. The Maximal model box is useful if you want to investigate several potential models, selecting or discarding explanatory terms as you search for the best model. This is discussed in detail in Section 1.11 of the *Guide to Regression, Nonlinear and Generalized Linear Models* in Genstat, which shows how setting the Maximal model allows the Change Model menu to be customized to make the search more convenient. Here, though, we have a single explanatory variate, and so there is no need to specify a maximal model. Clicking on Run produces the output below.

Regression analysis

Response variate: sales
Distribution: Poisson
Link function: Log
Fitted terms: Constant, temperature

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	1	52.61	52.614	52.61	<.001
Residual	28	32.05	1.145		
Total	29	84.66	2.919		

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

Message: the following units have large standardized residuals.

Unit	Response	Residual
30	137.00	2.87

Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant	4.3410	0.0303	143.49	<.001	76.78
temperature	0.01602	0.00222	7.22	<.001	1.016

Message: s.e.s are based on dispersion parameter with value 1.

The initial description contains the extra information that the data have a Poisson distribution, and that the *link* function (the transformation required to give a scale on which the model is linear) is the logarithm to base e. These are the two aspects required to characterize a generalized linear model. In the Log-linear modelling menu they are set automatically, but you can also select General model in the Analysis field to obtain a menu where you can set these explicitly, and thus fit any of Genstat's generalized linear models.

With generalized linear models, the summary of analysis contains *deviances* instead of sums of squares. Under the null hypothesis they have χ^2 distributions, and a quick rule-of-thumb is that their expected values are equal to their degrees of freedom.

However, some sets of data show *over-dispersion*. The residual deviance is then noticeably greater than its expectation and, instead of assessing the regression line by comparing its deviance with χ^2, you should use the deviance ratio (and assess this using an F distribution). You should also estimate the *dispersion parameter*, by choosing Estimate from the Dispersion Parameter radio buttons on either the Generalized Linear Model Options menu or the Generalized Linear Models Further Output menu (Figure 5.25). Genstat will then adjust the standard errors of the parameter estimates to take account of the over dispersion.

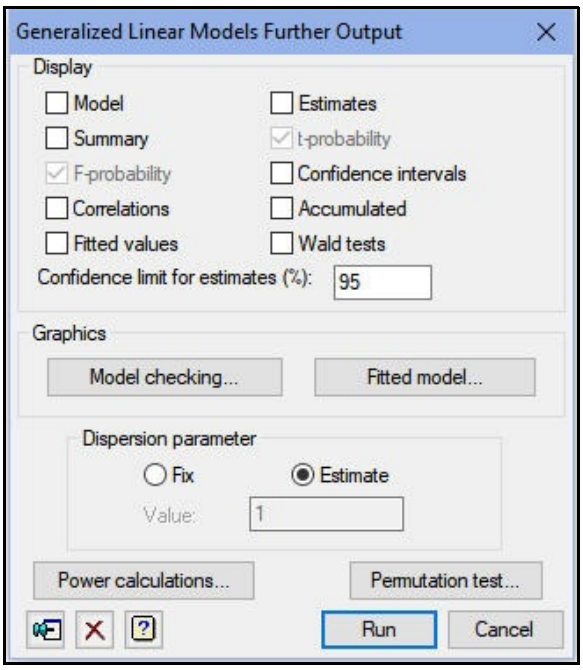


Figure 5.25

Note, however, that the residual deviance may be large not because of over dispersion, but simply because some important terms have been omitted from the model (and these may not even be available in the data set). You should then keep the dispersion parameter at the default value of 1, and continue to assess the deviances using χ^2 distributions. Further details are given in Section 3.5.1 of Part 2 of the *Guide to the Genstat Command Language*.

Here, though, the residual deviance is not substantially more than its expectation (as illustrated by the fact that its mean deviance is 1.145). So we can treat the regression deviance as χ^2 on one degree of freedom – and note that there seems to be a very strong effect of temperature on sales.

The fitted model can be displayed by clicking on the Further Output button to obtain the Generalized Linear Models Further Output menu (Figure 5.25), and then clicking on Fitted model to obtain the Graph of Fitted Model menu (Figure 5.26). This menu appears whenever you ask to plot the fitted model from one of the regressions menus where you yourself specify model to fit. There may then be several variates or factors to use for the x-axis or to define groups. Here there is only the variate `temperature`, so we enter that as the explanatory variable, and click on Run to plot the graph. We have chosen to plot the response on the linear predictor (i.e. the logarithmic) scale so that we can include 95% confidence limits for the response; see Figure 5.27. The line should be straight, so this also allows us to assess any nonlinearity in the response. The alternative is to plot with the y-axis on the natural scale.

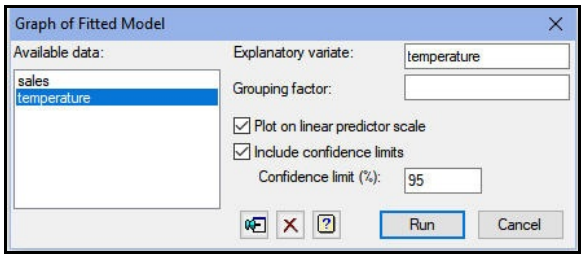


Figure 5.26

The y-axis of the fitted line in the graph (Figure 5.27) illustrates the logarithmic link transformation, and you can see the point with the large residual (on the top right of the plot). You can also produce the model-checking plots in the same way as in earlier sections.

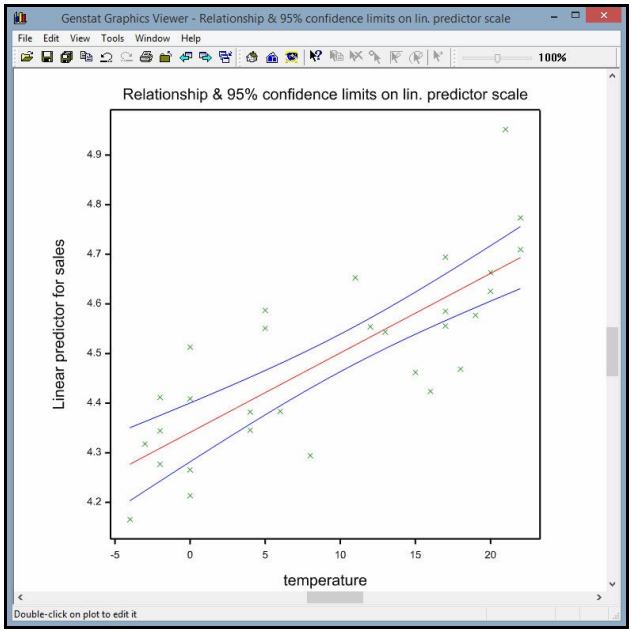


Figure 5.27

The Generalized Linear Models menu also has customized menus for binomial data, where each data value records a number of subjects responding out of a total number observed. The models will often involve factors as well as variates. Section 6.7 describes how to define the model formulae that you will then need to specify in the Maximal model and Model to be fitted boxes. Example analyses are described in Section 3.5 of Part 2 of the *Guide to the Genstat Command Language*.

5.8 Practical

The first column of file `Weed.GSH` contains counts of the numbers of weeds found growing in 10 plots. The second column records the amount of herbicide applied to each plot earlier in the year. Fit a log-linear model to see how the numbers of weeds relates to herbicide.

5.9 Regression commands

The commands for regression analysis give you more control over the fitting of models, and allow more complex models to be fitted as well. We describe here only those commands that carry out analyses like those already done in this chapter; for more information, see Chapter 3 of Part 2 of the *Guide to the Genstat Command Language*.

The `MODEL` directive must be used before any regression analysis, to specify the response variate; for example:

```
MODEL Pressure
```

`MODEL` can also define the distribution and link function of a generalized linear model (Section 5.7) using its `DISTRIBUTION` and `LINK` options.

A simple linear regression can then be fitted with the `FIT` directive:

```
FIT Age
```

The `FIT` directive has a `PRINT` option to control the output that is produced, so you could ask for all sections of output with the command:

```
FIT [PRINT=model,summary,estimates,correlations, \
  fitted,accumulated,confidence] Age
```

Alternatively, after fitting a model you can use the RDISPLAY directive to display further sections of output without refitting the model; it has a PRINT option just like FIT.

The RGRAPH procedure allows you to draw a picture of the fitted model. For example,

```
RGRAPH
```

draws a graph of a simple linear regression. After multiple regression, you can specify the explanatory variate or a grouping factor or both, as in

```
RGRAPH LogSulphur; GROUPS=Rain
```

The RCHECK procedure provides model checking. It has two parameters: the first specifies what to display in the graph (residuals, Cook's statistics or leverages) and the second specifies the type of graph (composite, histogram, fittedvalues, index, Normal or halfNormal). For example,

```
RCHECK
```

draws the composite picture, or you could plot just the residuals against the fitted-values by

```
RCHECK residual; fitted
```

The permutation tests are provided by the RPERMTEST procedure. Those in 5.1 were produced by the statement

```
RPERMTEST [PRINT=probability; CONSTANT=estimate; SEED=0;\
  NTIMES=4999] Age
```

The RKEEP directive allows you to extract information into standard structures. It has many parameters, for each piece of information. To save the residuals and fitted values, for example, you can give a command like the following after fitting the model:

```
RKEEP RESIDUALS=resids; FITTEDVALUED=fvalues
```

Multiple linear regressions can be fitted simply by listing the explanatory variates in the FIT directive. The list may also include factors, which allows you to fit simple or multiple linear regression models with groups; for example:

```
FIT Windsp,Rain
```

The parameter of FIT can also be a model formula (defined in Section 6.7), which can include interactions between factors or variates or both; for example:

```
FIT Windsp*Rain
```

fits the linear effect of the variate Windsp, the main effect of the factor Rain, and the interaction between them (which represents separate linear effects of Windsp for each level of Rain).

Standard curves are fitted using the FITCURVE directive, which has a parameter and a PRINT option just like FIT, together with a CURVE option to choose the type of curve; for example:

```
FITCURVE [PRINT=summary; CURVE=exponential] Nitrogen
```

5.10 Other facilities

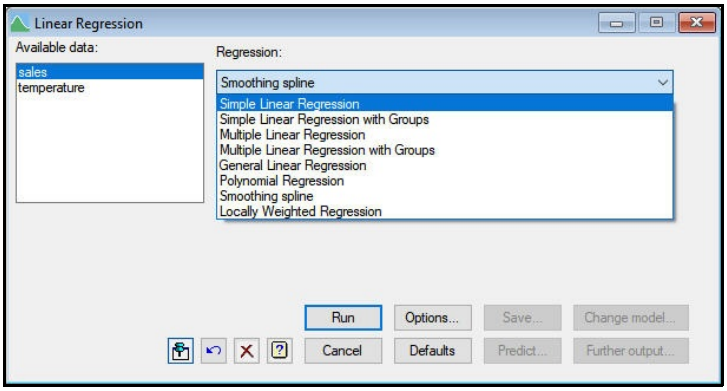


Figure 5.28

In addition to the simple linear regression models (with and without groups), described in Section 5.1 and 5.3, Figure 5.28 shows that the Linear Regression menu covers several other types of model, including multiple linear regression (with or without groups), polynomial regression, smoothing splines and locally weighted regression. It also has a General linear regression setting which allows you to explore and fit general regression models defined by model formulae involving arbitrary collections of explanatory variates and factors.

Other menus in the regression option, but not described in this Chapter, include ordinal regression, all-subsets regression, screening tests, split-line regression, generalized linear mixed models, hierarchical generalized linear models, regression trees and quantile regression. The repeated-measurements option also has menus to fit linear regressions and standard curves when the residuals follow an auto-regressive or power-distance correlation model.

Many of these are described in the *Guide to Regression, Nonlinear and Generalized Linear Models in Genstat*, which can be accessed from within Genstat *for Windows* by selecting sub-options of the Genstat Guides option of the Help menu on the menu bar (see Figure 1.6). Information on the others can be obtained by clicking on the Help buttons on the relevant menus.

6 Analysis of variance

Genstat has very comprehensive facilities for analysis of variance. Almost all of these can be accessed using custom menus. In this chapter, we start with the simplest design, a one-way completely randomized experiment, before introducing factorial experiments, which have more than one *treatment* or *fixed effect*. We use an experiment with a randomized block design to show how to deal with *blocks*, which involve more than one *stratum* or *source of error* in the analysis, and extend this idea by analysing a split-plot design.

A more comprehensive guide to the analysis of variance menus can be found in the *Guide to ANOVA and Design in Genstat*, while further details of the commands and the underlying statistical theory are in the *Guide to the Genstat Command Language, Part 2 Statistics*, Chapter 4. These can both be accessed from within Genstat *for Windows* by selecting sub-options of the Genstat Guides option of the Help menu on the menu bar (see Figure 1.6).

6.1 One-way analysis of variance

Diet	Weight
a	81.5 80.7 80.3 79.8
b	81.6 81.9 80.4 80.4
c	83.5 81.6 82.2 81.3
d	82.4 83.1 82.8 81.8
e	83.2 82.8 82.1 82.1

We shall start with a simple one-way analysis of variance. This experiment was set up to study the effect of a dietary supplement on the gain in weight of rats. There were five different treatments (representing different amounts of the supplement) and 20 rats were allocated at random, four to each treatment.

The data are available in spreadsheet file `Rat.gsh` (Figure 6.1) which can be opened from within Genstat using the Example Data Sets menu, as explained in Section 4.1. There are two columns of data: the name *`diet`* is in italics, showing that this column is a factor, and `weight` is a variate.

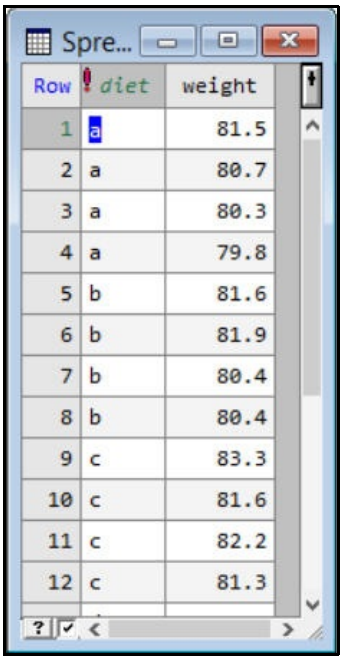


Figure 6.1

The Analysis of Variance option of the Stats menu on the menu bar (Figure 6.2) offers four sub-options. In this chapter we will first use the simple menu for one- and two-way analysis of variance. Then in Sections 6.5 and 6.8 we introduce the general Analysis of Variance menu, which accesses the full power of Genstat's analysis of variance facilities. The menu for Unbalanced ANOVA, opened by the Unbalanced Designs sub-option, is described in Section 7.6 of the *Guide to ANOVA and Design in Genstat*.

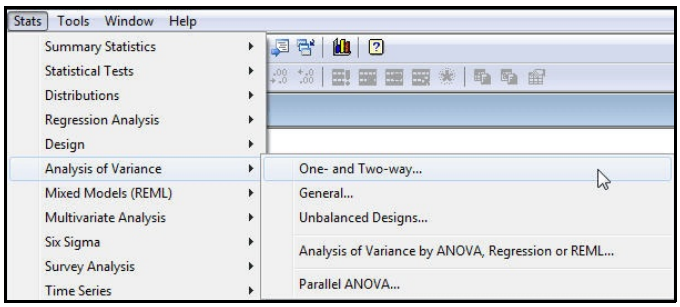


Figure 6.2

The Design radio buttons, at the top, customize the menu for one- or two-way analyses. Figure 6.3 shows the form of the menu when the One-way button is pressed. We enter the name of the y-variate `weight` into the Y-variate window, and the name of the treatment factor `diet` into the Treatments window. As with the boxes in other Genstat menus, the relevant identifiers can be selected from the Available data window. Clicking on Run then produces the analysis shown below.

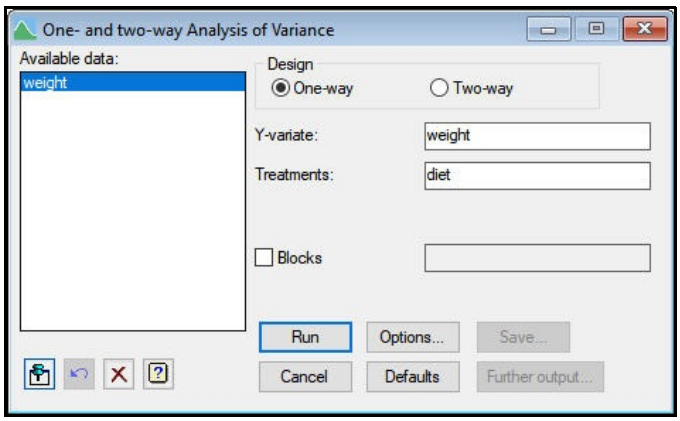


Figure 6.3

Analysis of variance

Variate: weight

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
diet	4	12.7930	3.1982	6.32	0.003
Residual	15	7.5925	0.5062		
Total	19	20.3855			

Information summary

All terms orthogonal, none aliased.

Tables of means

Variate: weight

Grand mean 81.76

diet	a	b	c	d	e
	80.58	81.08	82.10	82.53	82.55

Standard errors of differences of means

Table	diet
rep.	4
d.f.	15
s.e.d.	0.503

Here we have just the default output. This contains an analysis-of-variance table, in the standard format, followed by an "information summary" which tells us that this is a very straightforward analysis with no complications. It then shows the grand (or overall) mean, and a table of means for the different diets with an accompanying standard error to assess differences between pairs of diet means. Standard errors of means and least significant differences can be obtained by modifying the ANOVA Options menu shown in Figure 6.6. We also show, later in this chapter, how to obtain further output, including plots of means and residuals.

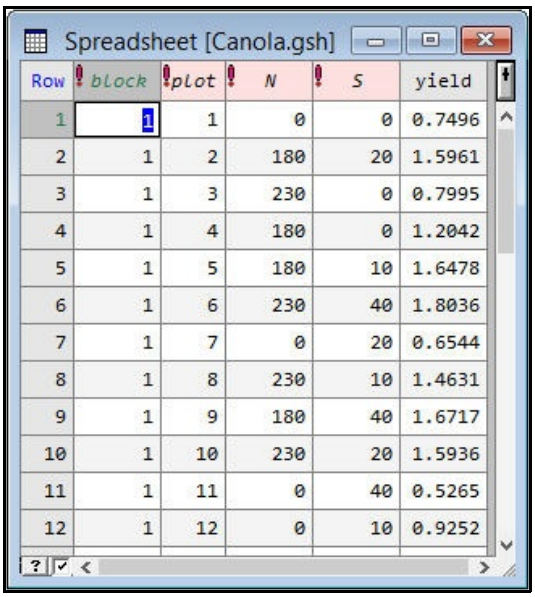
6.2 Practical

Spreadsheet `Octane.gsh` contains data from an experiment to study the effect of different additives on the octane level of gasoline (P.W.M. John, Statistical Design and Analysis of Experiments, page 46). There were 5 types of gasoline (A-E), and 4 observations on each. Use analysis of variance to assess whether there are differences in octane level between the gasolines.

6.3 Two-way analysis of variance

We now consider a more complicated example. This is a field experiment performed to examine the effects of sulphur and nitrogen fertilizers on the yield of canola. So there are two treatment factors, which we shall call `S` and `N`. The experiment used a randomized-block design, so there is also a factor, here called `block`, to indicate the block to which each of the experimental plots belonged.

To analyse the experiment we first load the data from the file `Canola.gsh`, producing the spreadsheet in Figure 6.4.



Row	block	plot	N	S	yield
1	1	1	0	0	0.7496
2	1	2	180	20	1.5961
3	1	3	230	0	0.7995
4	1	4	180	0	1.2042
5	1	5	180	10	1.6478
6	1	6	230	40	1.8036
7	1	7	0	20	0.6544
8	1	8	230	10	1.4631
9	1	9	180	40	1.6717
10	1	10	230	20	1.5936
11	1	11	0	40	0.5265
12	1	12	0	10	0.9252

Figure 6.4

Initially, we shall ignore the (randomized-block) structure of the design, and use the data merely to illustrate how to perform a two-way analysis of variance. Pressing the Two-way button generates the menu shown in Figure 6.5. The y-variate `yield` is entered into the Y-variate box as before and there are now two boxes, Treatment 1 and Treatment 2, into which the two treatment factors (`N` and `S`) are entered. The Include interaction box allows you to decide whether you want to fit the interaction between the factors. Here we have checked the box so that the interaction between nitrogen and sulphur will be included.

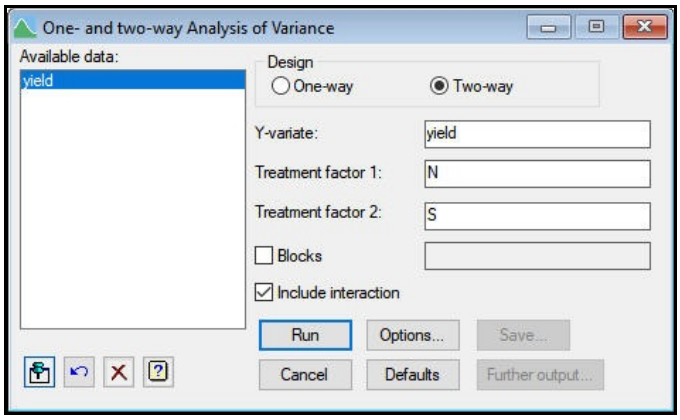


Figure 6.5

Figure 6.6 shows the ANOVA Options menu for one- and two-way analysis of variance (selected by clicking the Options button). This allows you to control the output produced initially from the analysis. The menu consists of a collection of boxes that can be checked to select the output components that you want. By default the following are selected: AOV Table (analysis-of-variance table), Information (details of any large residuals, non-orthogonality or aliasing in the model), Means (tables of means), and F-probability (probabilities for variance ratios in the analysis-of-variance table). Notice that Genstat can allow the use of multiple comparisons between means. The appearance of the Multiple comparisons button is controlled by a check box in the Menus page of the Options menu; see Section 1.7.

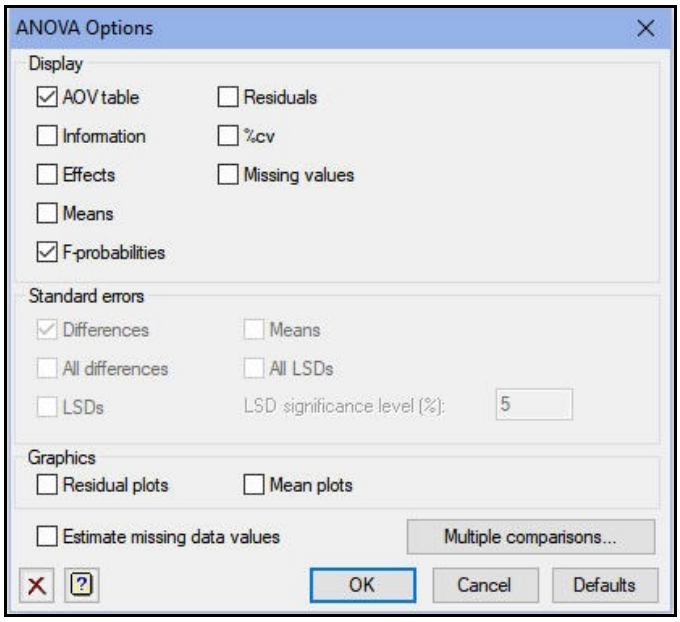


Figure 6.6

If we clear the Information and Means boxes , as in Figure 6.6, and click OK in the ANOVA Options menu followed by Run in the One- and two-way Analysis of Variance menu, only the analysis-of-variance table will be produced, as shown below.

Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
N	2	4.59223	2.29611	42.56	<.001
S	3	0.97720	0.32573	6.04	0.003
N.S	6	0.64851	0.10808	2.00	0.105
Residual	24	1.29476	0.05395		
Total	35	7.51269			

The table now has lines for three *treatment terms*: `N` represents the main effect of nitrogen, that is the overall way in which yield responds to nitrogen. Similarly `S` represents the main effect of sulphur, while `N.S` represents the interaction between nitrogen and sulphur. The interaction assesses the way in which the effect of nitrogen on yield differs according to the amount of sulphur or, equivalently, the way in which the sulphur effect differs according to the amount of nitrogen. If there is no interaction, we could decide on the best amount of nitrogen to apply without needing to consider how much sulphur will be used (and how much sulphur to use without needing to think about the amount of nitrogen).

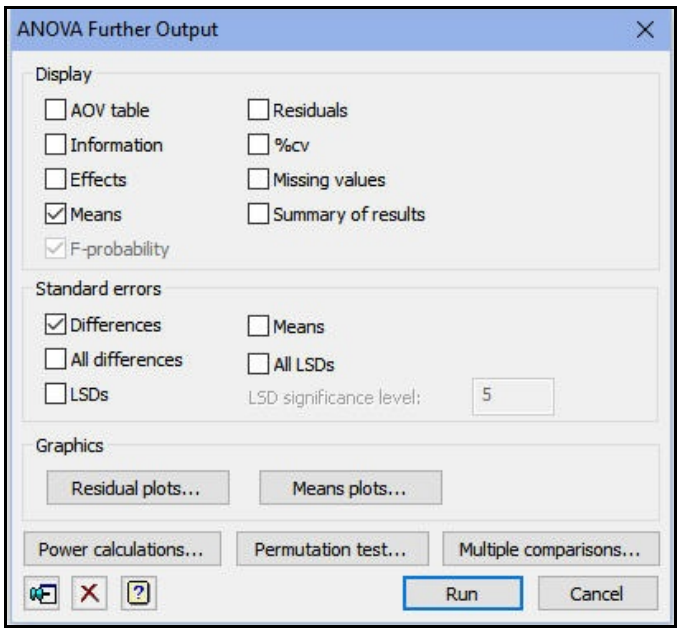


Figure 6.7

The Further output button in the Analysis of Variance menu allows additional analysis of variance output to be obtained. Many of the components, shown in Figure 6.7, are the same as those in the Options menu. So we can obtain tables of means by checking the Means box, and then clicking the Run button.

Tables of means

Variate: yield

Grand mean 1.104

N	0	180	230
	0.601	1.313	1.398

S	0	10	20	40
	0.829	1.155	1.167	1.266

N	S	0	10	20	40
0		0.560	0.770	0.524	0.552
180		0.894	1.289	1.525	1.545
230		1.032	1.404	1.454	1.700

Standard errors of differences of means

Table	N	S	N S
rep.	12	9	3
d.f.	24	24	24
s.e.d.	0.0948	0.1095	0.1896

Notice that Genstat has produced a table of means for every term in the analysis of variance, each with an appropriate standard error for assessing differences between pairs of means. The measures of variability to accompany the means (standard errors of difference, least significant differences or standard errors of means) are controlled by the Standard errors check boxes. You can click more than one of these, for example to have least significant differences as well as standard errors of differences. If you do request least significant differences, a further box appears allowing you to change the significance level used in their calculation from the default of 5%.

6.4 Practical

Spreadsheet file `Ratfactorial.gsh` contains data from an experiment to study the effect of 6 different diets on the gain in weight of rats (data from Snedecor and Cochran, Statistical Methods p.305). Each diet was at either High or Low protein (factor `Amount`), and the protein was derived from either Beef, Cereal or Pork (factor `Source`).

Analyse the data as a 3×2 factorial, and assess whether there is evidence for an interaction between `Amount` and `Source`.

6.5 Randomized-block designs

The randomized-block design is perhaps the simplest type of designed experiment. In these designs, the experimental units are grouped together into sets known as *blocks* with the aim that units in the same block will be more similar than units in different blocks. Each block contains the same number of replicates of each treatment combination (usually one of each), and the allocation of the treatments is randomized independently within each block. In the analysis, the aim is to estimate and remove the between-block differences so that the treatment effects can be estimated more precisely. In our example, there is a factor called `block` to indicate the "block" of land to which each plot belonged. In other examples the blocking factor might represent different litters of animals, or different days on which the experiment was conducted, and so on.

You can extend the One- and two-way Analysis of Variance menu to allow for blocking simply by checking the Blocks box; see Figure 6.8. A window then appears into which you should enter the name of the block factor (here `block`).

Clicking the Run button generates the analysis-of-variance table again.

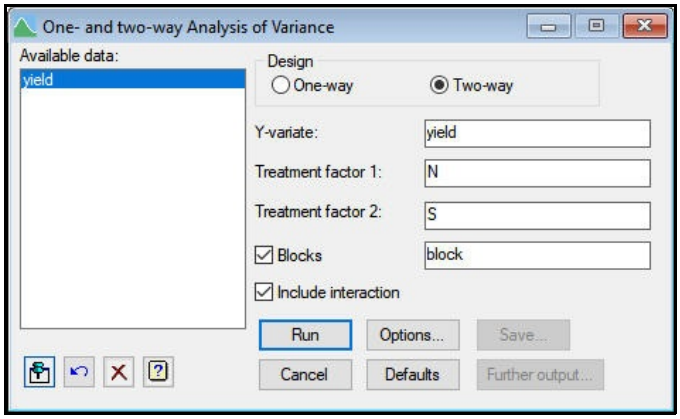


Figure 6.8

Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
block stratum	2	0.30850	0.15425	3.44	
block.*Units* stratum					
N	2	4.59223	2.29611	51.22	<.001
S	3	0.97720	0.32573	7.27	0.001
N.S	6	0.64851	0.10808	2.41	0.061
Residual	22	0.98625	0.04483		
Total	35	7.51269			

The differences between the blocks are placed into the line entitled "`block` stratum", while the "`block.*Units*` stratum" contains the variation of the plots within blocks. The variance ratio for the `block` stratum compares the variability of the blocks of land with the variability of the individual plots within each block – and its value of 3.44 shows that it was worthwhile using the design in this experiment. This can be confirmed also by the fact that the mean square for the Residual has decreased from 0.054 to 0.045. (The Residual line now represents the random variability of the experimental plots after removing block differences as well as the effects of the treatments.) So the standard errors of differences of means (printed using the ANOVA Further Output menu as before) will also be smaller.

Tables of means

Variate: yield

Grand mean 1.104

N	0	180	230
	0.601	1.313	1.398

S	0	10	20	40
	0.829	1.155	1.167	1.266

N	S	0	10	20	40
0		0.560	0.770	0.524	0.552
180		0.894	1.289	1.525	1.545
230		1.032	1.404	1.454	1.700

Standard errors of differences of means

Table	N	S	N S
rep.	12	9	3
d.f.	22	22	22
s.e.d.	0.0864	0.0998	0.1729

Now that we have performed an analysis we can click the Save button to obtain the ANOVA Save Options menu, allowing us to save variates of residuals and fitted values, and tables of means. After checking the appropriate box, a window (entitled In:) will appear into which you enter the identifier of the structure in which the information is to be saved. Figure 6.9 saves the residuals in a variate called `yieldres` and an `N` by `S` table of means in a table called `NSmeans`. You can save means for any of the treatment terms in the analysis; the name of the term is selected from the list in the Treatment term box. By checking the Display in spreadsheet box, we can arrange for the table of means to be loaded automatically into a table spreadsheet, from which it can conveniently be transferred into other documents, as explained in Section 4.8.

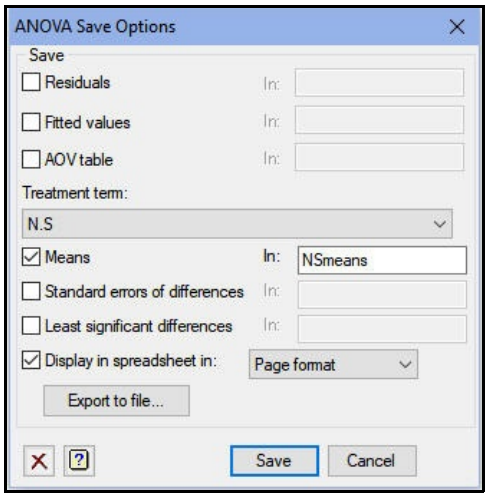


Figure 6.9

To save other information from `ANOVA`, such as sums of squares, degrees of freedom and so on, you need to enter command mode and use the directive `AKEEP` (see Section 6.10).

We can also obtain plots of residuals and of means. Clicking the Residual plots button in the ANOVA Further Output menu produces the ANOVA Residual Plots menu as shown in Figure 6.10. If you check the Histogram box, a histogram is plotted of the residuals. The Fitted values box produces a plot of residuals against fitted values. Normal produces a Normal plot and Half Normal a half-Normal plot of the residuals. Here we leave the default settings (shown in Figure 6.10) and generate the output shown in Figure 6.11.

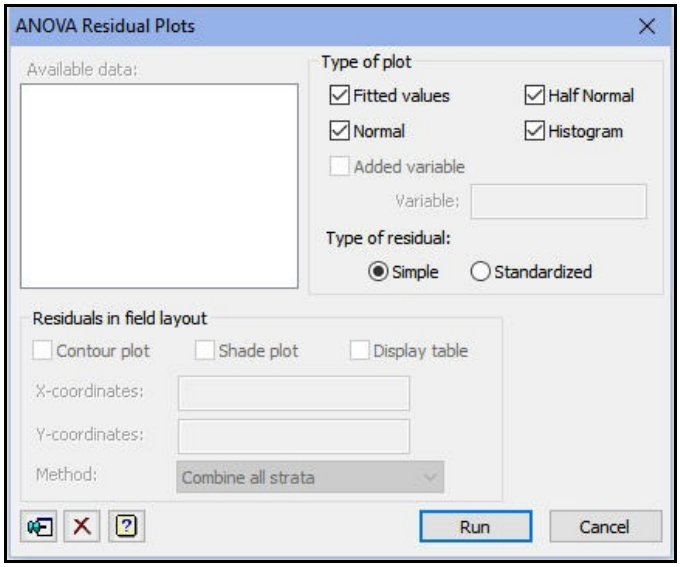


Figure 6.10

These plots are used in exactly the same way as the regression residual plots (Section 5.1). As in regression, the analysis is robust to small departures from Normality. If the residuals do not seem to follow a Normal distribution, however, you could use a permutation test as an alternative to the F tests in the analysis of variance table.

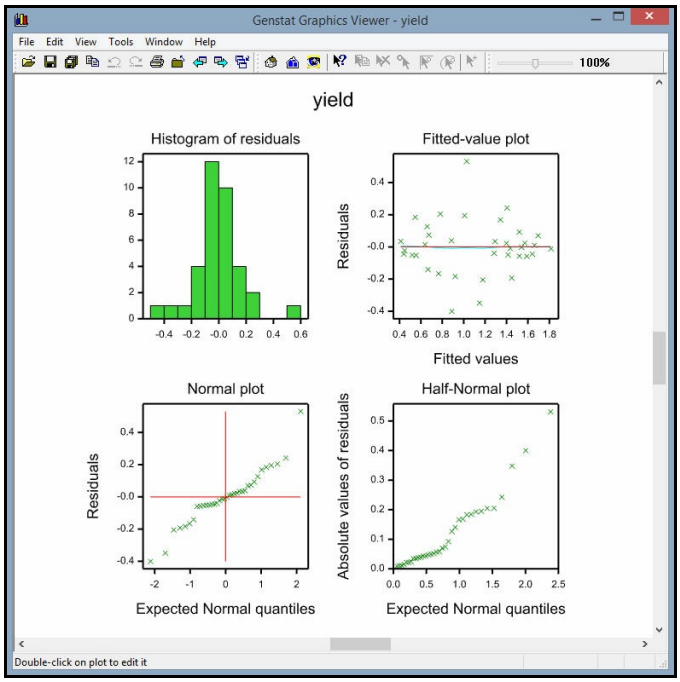


Figure 6.11

Clicking on the Permutation test button in the ANOVA Further Output menu (Figure 6.7) produces the menu in Figure 6.12. This asks Genstat to make 4999 random permutations (see the Number of permutations box), repeating the analysis of variance each time. The value 0 in the Seed box means that Genstat will select a seed automatically for the random numbers that are used to generate the permutations. If you have used random numbers already in the current run of Genstat, the seed will be chosen to continue the existing sequence. Otherwise it is initialized automatically (and the value is printed in the output).

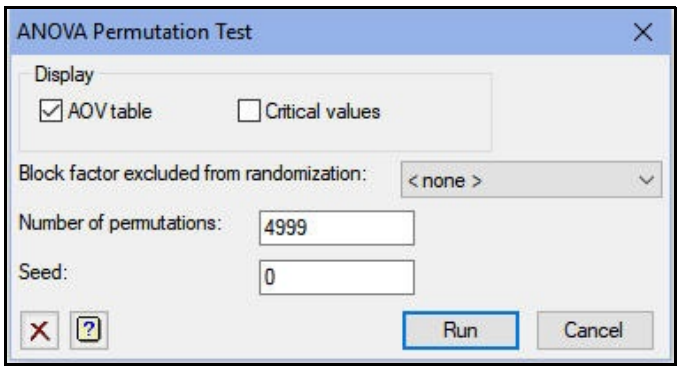


Figure 6.12

The random permutations use the same method of randomization as would have been used in the original design. So, with a randomized-block design, the allocation of the treatments is (re)randomized independently within each block. The Block factor excluded from randomization box is relevant with specialized designs, such as cross-over designs, where some blocking factors must be excluded from the randomization.

The probabilities for the variance ratios are now determined from their distributions over the randomly permuted data sets. For example, below, N has a probability <.001, which means that the observed data set had one of the 5 largest variance ratios for N out of the 5000 sets that have been examined (1 observed data set + 4999 randomly permuted data sets).

Message: Default seed for random number generator used with value 299759

Analysis of variance

Variate: yield
Probabilities determined from 4999 random permutations

Source of variation	d.f.	s.s.	m.s.	v.r.	prob.
block stratum	2	0.3085	0.1543	3.44	
block.*Units* stratum					
N	2	4.5922	2.2961	51.22	<.001
S	3	0.9772	0.3257	7.27	0.001
N.S	6	0.6485	0.1081	2.41	0.067
Residual	22	0.9863	0.0448		
Total	35	7.5127			

If you ask for more permutations than the number that are possible for your data, Genstat will instead do an *exact test*, which uses each permutation once.

As described already in Section 1.1, you can click the Genstat icon on the task bar to return to Genstat and the Analysis of Variance menu. If you again click the Further output button and then click the Mean plots button in the ANOVA Further Output menu, the ANOVA Means Plots menu appears (Figure 6.13). This menu plots tables of means from the analysis. The Method box contains option buttons to select the type of plot. Means represents each mean by a point, Lines plots the point at the means and draws lines between them, Data draws just the lines together with the original data values, and Barchart plots the means in a barchart. The Standard error bar box allows you to choose how to represent the variability of the means.

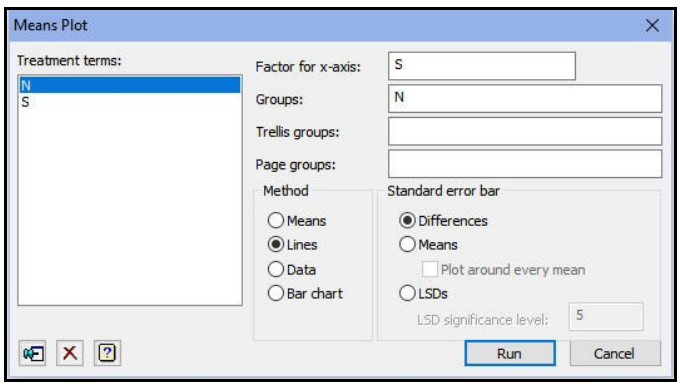


Figure 6.13

The Factor for x-axis is the factor against whose levels the means are to be plotted, while Groups specifies the other factor when you want to plot a two-way table in a single window. Separate lines are drawn for the groups, and the points in each group are plotted using different pens. Alternatively, if you set Trellis groups, the lines are drawn in separate windows (one for each group) in a "trellis" arrangement. If neither Factor for x-axis, nor Groups. nor Trellis groups, are specified, the first two-way table of means in the analysis is plotted in a single window, or for the first one-way table if there were no two-way tables. Here we have set Method to Lines, the Factor for x-axis to S, Groups to N, and selected Differences for the Standard error bar. The resulting graph is shown in Figure 6.14.

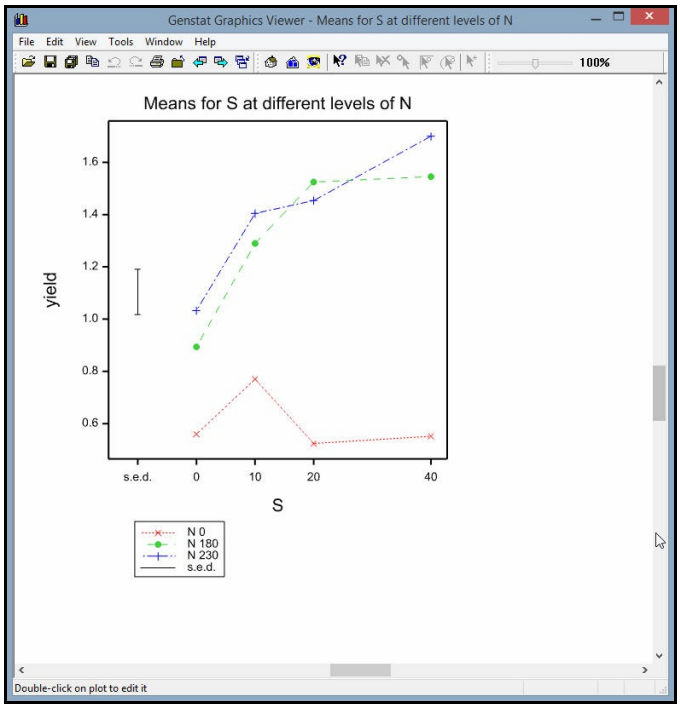


Figure 6.14

You can also analyse one- and two-way designs using the general Analysis of Variance menu, obtained by clicking on the General line in Figure 6.2. In this menu, the type of design is selected using the Design list box. The possibilities range from simple One-way ANOVA to General analysis of variance – each with its appropriate boxes and buttons. Figure 6.15 shows the Two-way ANOVA (in randomized blocks) option, set up to repeat the analysis that we have already produced using the One- and two-way Analysis of Variance menu. As well as offering further types of design, the menu also provides more sophisticated types of analysis including the fitting of contrasts; see Section 3.2 of the *Guide to ANOVA and Design in Genstat*.

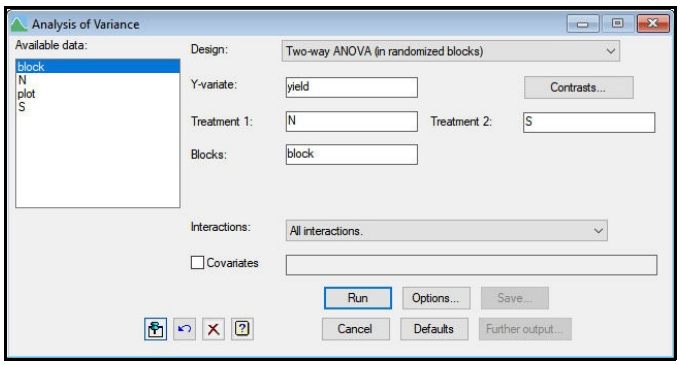


Figure 6.15

6.6 Practical

Seven litters each of five rats were used in a randomized-block design (with litters as blocks) to study the effects of different diets on the gain in weight of rats. Analyse the data, in file `Ratblocks.gsh`, to see whether there are any differences between the diets.

```
"litter  diet    gain"
      1     B    87.9
      1     D    80.0
      1     A    76.4
      1     C    54.6
      1     E    76.2
      2     C    51.7
      2     A    74.6
      2     E    78.9
      2     D    63.2
      2     B    85.6
      3     D    70.8
      3     C    62.2
      3     A    77.6
      3     B    88.6
      3     E    83.2
      4     A    83.0
      4     E    70.7
      4     C    80.6
      4     D    84.9
      4     B   103.6
      5     C    83.7
      5     B   100.6
      5     E   101.3
      5     A    94.5
      5     D    76.6
      6     E    71.9
      6     D    72.9
      6     B    54.2
      6     C    47.4
      6     A    55.8
      7     A    54.9
      7     C    76.8
      7     D    68.6
      7     E    82.8
      7     B    65.7
```

6.7 Syntax of model formulae

The structure of the design and the treatment terms to be fitted in a Genstat analysis of variance are specified by *model formulae*. In the simpler menus, like those we have used earlier in this chapter, the formulae are constructed automatically behind the scenes. However, for the more advanced menus and analyses you will need to specify your own formulae.

Several of the menus allow you to specify any number of treatment factors, interactions and so on. So, for example, the General analysis of variance, the General treatment structure (no blocking) and the General treatment structure (in randomized blocks) menus all have a box entitled Treatment structure into which a formula (known as the *treatment formula*) needs to be entered.

The General Analysis of Variance menu also allows you to define any *underlying structure* for the design (for example completely randomized, randomized-block, split-plot, split-split-plot, and so on). This is specified by a model formula (the *block formula*) which is entered into the Block structure box; this can be left blank with unstructured (completely randomized) designs. This formula defines the strata and thus the error terms for the analysis.

In its simplest form, a model formula is a list of *model terms*, linked by the operator "+". For example,

```
A + B
```

is a formula containing two terms, A and B, representing the main effects of factors A and B respectively. *Higher-order terms* (like interactions) are specified as series of factors separated by dots, but their precise meaning depends on which other terms the formula contains, as we explain below. The other operators provide ways of specifying a formula more succinctly, and of representing its structure more clearly.

The *crossing operator* * is used to specify factorial structures. The formula

```
N * S
```

was used by Genstat to specify the two-way analysis of variance introduced in Section 6.3. This is expanded to become the formula

```
N + S + N.S
```

which has three terms: N for the nitrogen main effect, S for the main effect of sulphur, and N.S for the nitrogen by sulphur interaction. Higher-order terms like N.S represent all the joint effects of the factors N and S that have not been removed by earlier terms in the formula. Thus here it represents the interaction between nitrogen and sulphur as both main effects have been removed.

The other most-commonly used operator is the *nesting operator* (/). This occurs most often in block formulae. For example, the formula

```
block /  plot
```

is expanded to become the formula

```
block + block.plot
```

As the formula contains no "main effect" for plot, the term block.plot would represent *plot-within-block* effects (that is the differences between individual plots after

removing any overall similarity between plots that belong to the same block). This is similar to the block model for the randomized design in Section 6.5 except that we have the factor `plot` instead of `*Units*`.

A formula can contain more than one of these operators. The three-factor factorial model

```
A * B * C
```

becomes

```
A + B + C + A.B + A.C + B.C + A.B.C
```

and the nested structure

```
block / wplot / subplot
```

which occurs as the block model of a split-plot design (Section 6.8) becomes

```
block + block.wplot + block.wplot.subplot
```

They can also be mixed in the same formula. For example, the factorial-plus-added-control study in Section 3.5 of the *Guide to ANOVA and Design in Genstat* has treatment structure

```
Control / (Drug * Dose)
```

which expands to

```
Control + Control.Drug + Control.Dose + Control.Drug.Dose
```

In general, if *l* and *m* are two model formulae:

```
l * m  =  l + m + l.m
l / m  =  l + fac(l).m
```

(where *l*`.`*m* is the sum of all pairwise dot products of a term in *l* and a term in *m*, and `fac(`*l*`)` is the dot product of all factors in *l*). For example:

```
(A + B) * (C + D) = (A + B) + (C + D) + (A + B).(C + D)
                  = A + B + C + D + A.C + A.D + B.C + B.D
(A + B)/C = A + B + fac(A + B).C  =  A + B + A.B.C
```

Terms in the treatment formula can be partitioned into contrasts by specifying a function of the factor.

`COMPARISON(`*factor*`;` *scalar*`;` *matrix*`)` partitions the *factor* into the comparisons specified by the *matrix*. There is a row of the matrix for each comparison, and the *scalar* specifies how many of them are to be fitted.

`POL(`*factor*`;` *scalar*`;` *variate*`)` partitions the *factor* into polynomial contrasts (linear, quadratic and so on). The *scalar* gives the maximum order of contrast (1 for linear only, 2 for linear and quadratic, and so on) and the *variate* gives a numerical value for each level of the factor. If the variate is omitted, the levels defined when the factor was declared will be used.

`REG(`*factor*`;` *scalar*`;` *matrix*`)` partitions the *factor* into the (user-defined) regression contrasts specified by the coefficients in each row of the *matrix*. The *scalar* defines the number of contrasts to be fitted.

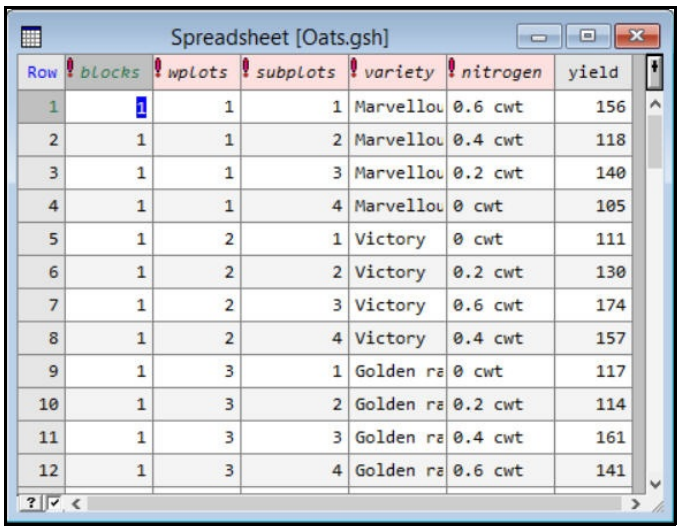
6.8 Split-plot designs

V3 N3	V3 N2	V3 N2	V3 N3
V3 N1	V3 N0	V3 N0	V3 N1
V1 N0	V1 N1	V2 N0	V2 N2
V1 N3	V1 N2	V2 N3	V2 N1
V2 N0	V2 N1	V1 N1	V1 N2
V2 N2	V2 N3	V1 N3	V1 N0
V3 N2	V3 N0	V2 N3	V2 N0
V3 N1	V3 N3	V2 N2	V2 N1
V1 N3	V1 N0	V1 N2	V1 N3
V1 N1	V1 N2	V1 N0	V1 N1
V2 N1	V2 N0	V3 N2	V3 N3
V2 N2	V2 N3	V3 N1	V3 N0
V2 N1	V2 N2	V1 N2	V1 N0
V2 N3	V2 N0	V1 N3	V1 N1
V3 N3	V3 N1	V2 N3	V2 N2
V3 N2	V3 N0	V2 N0	V2 N1
V1 N0	V1 N3	V3 N0	V3 N1
V1 N1	V1 N2	V3 N2	V3 N3

We now show how to analyse split-plot designs with the analysis-of-variance menus. These designs were devised originally for agricultural experiments where some of the factors can be applied to smaller plots of land than others. Here there are two treatment factors: three different varieties of oats (labelled V1, V2 and V3 on the plan), and four levels of nitrogen (labelled N0 to N3). Because of limitations on the machines for sowing seed, different varieties cannot conveniently be applied to plots as small as those that can be used for the different rates of fertilizer. So the design was set up in two stages. First of all, the blocks were each divided into three plots of the size required for the varieties, and the three varieties were randomly allocated to the plots within each block (exactly as in the randomized blocks design). Then each of these plots, or *whole-plots* as they are usually known, was split into four *sub-plots* (one for each rate of nitrogen), and the allocation of nitrogen was randomized independently within each whole-plot.

Split-plot designs occur not only in field experiments, but also in animal trials (where, for example, the same diet may need to be fed to all the animals in a pen but other treatments may be applied to individual animals), or in industrial experiments (where different processes may require different sized batches of material), or even in cookery experiments. There can also be more than one treatment factor applied to any size of unit.

Figure 6.16 shows Genstat spreadsheet, `Oats.gsh`, which contains the data from the experiment. The `blocks` factor (column 1) indicates the block to which each of the individual experimental plots belongs, `wplots` (column 2) numbers the whole plots within each block and `subplots` (column 3) numbers the sub-plots within each whole plot. The fourth and fifth columns contain the treatment factors, `variety` and `nitrogen`, and final column is the y-variate `yield`.



Row	blocks	wplots	subplots	variety	nitrogen	yield
1	1	1	1	Marvellou	0.6 cwt	156
2	1	1	2	Marvellou	0.4 cwt	118
3	1	1	3	Marvellou	0.2 cwt	140
4	1	1	4	Marvellou	0 cwt	105
5	1	2	1	Victory	0 cwt	111
6	1	2	2	Victory	0.2 cwt	130
7	1	2	3	Victory	0.6 cwt	174
8	1	2	4	Victory	0.4 cwt	157
9	1	3	1	Golden ra	0 cwt	117
10	1	3	2	Golden ra	0.2 cwt	114
11	1	3	3	Golden ra	0.4 cwt	161
12	1	3	4	Golden ra	0.6 cwt	141

Figure 6.16

The data can be analysed by selecting the Split-plot design setting in the Design drop-down list box of the general Analysis of Variance menu, as shown in Figure 6.17. The factor defining the blocks is entered into the Blocks box, and the factor defining the whole-plots within each block is entered into the Whole-plots box. There is no need to specify a factor for the sub-plots but, if one is available, it can be entered into the Sub-plots box.

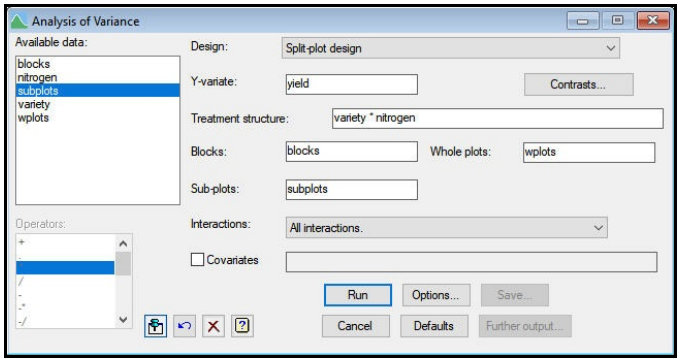


Figure 6.17

The treatment terms to be fitted are specified by entering a model formula (see Section 6.7) into the Treatment structure box. The factors for the formula can be selected from the Available data window, and the available operators are listed in the Operators window. Here we have specified

```
variety * nitrogen
```

to indicate that we want the main effects of variety and nitrogen, and their interaction. As explained in Section 6.7, Genstat expands the `*` operator so that the formula becomes

```
variety + nitrogen + variety.nitrogen
```

(and if you prefer you can enter this expanded form, where the terms to be fitted in the analysis are all specified explicitly, instead).

The Interactions box can be used to control the level of interactions to be fitted – you can indicate either All interactions, as here, or just main effects (No interactions), or select Specify level of interaction to specify the required level of interaction (that is, set a limit on the maximum number of factors in the treatment terms that are fitted).

As usual, the Y-variate box is set to the Genstat variate containing the data values (here `yield`). If we now click the Run button, after resetting the options to display AOV table, Information and Means, the output below is produced.

Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
blocks stratum	5	15875.3	3175.1	5.28	
blocks.wplots stratum					
variety	2	1786.4	893.2	1.49	0.272
Residual	10	6013.3	601.3	3.40	
blocks.wplots.subplots stratum					
nitrogen	3	20020.5	6673.5	37.69	<.001
variety.nitrogen	6	321.8	53.6	0.30	0.932
Residual	45	7968.8	177.1		
Total	71	51985.9			

Message: the following units have large residuals.

blocks 1	31.4	s.e.	14.8

Tables of means

Variate: yield

Grand mean 104.0

variety	Victory	Golden rain	Marvellous
	97.6	104.5	109.8

nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
	79.4	98.9	114.2	123.4

variety	nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
Victory		71.5	89.7	110.8	118.5
Golden rain		80.0	98.5	114.7	124.8
Marvellous		86.7	108.5	117.2	126.8

Standard errors of differences of means

Table	variety	nitrogen	variety nitrogen
rep.	24	18	6
s.e.d.	7.08	4.44	9.72
d.f.	10	45	30.23
Except when comparing means with the same level(s) of variety			7.68
d.f.			45

There are now three strata.

The `blocks` stratum contains the variation between the blocks. The blocks all contain exactly the same treatments (one of each of the possible combinations of variety and level of nitrogen), so none of this variation can arise from the effects of the treatments. There are hence no treatment terms estimated in this stratum. (For the same reason none of the treatment terms was estimated in the `block` stratum of the randomized-block design in Section 6.5.)

However, varieties (which were applied to complete whole-plots in the design), are estimated in the `blocks.wplots` stratum; in conventional terminology this is called the stratum for whole-plots within blocks and it contains the variation between the whole-plots after eliminating differences between the blocks. The variance ratio for varieties is calculated by dividing the `variety` mean square by the `blocks.wplots` residual mean square. It is easy to see that this is the correct thing to do. When we look to see whether the varieties differ we are really trying to answer the question: "Do the yields from the three sets of whole-plots, on the first of which the variety Victory was grown, on the second Golden rain, and on the third Marvellous, differ by more than the amount that we would expect for any three randomly chosen sets of whole-plots (each set containing one whole-plot from every block)?". Technically, variety is said to be *confounded* with whole plots.

The terms for `nitrogen`, which was applied to sub-plots, and for the `variety.nitrogen` interaction are both estimated in the stratum for sub-plots within whole-plots (`blocks.wplots.subplots`). Thus, these are both compared against the residual of that stratum, which measures the variability of the sub-plots after eliminating differences between the whole-plots (and blocks).

The standard errors accompanying the tables of means also take account of the stratum where each treatment term was estimated.

The `variety` s.e.d. of 7.08 = √(2×601.3/24) is based on the residual mean square for `blocks.wplots`, while that for `nitrogen` (4.44 = √(2×177.1/18)) is based on that for `blocks.wplots.subplots`. The `variety` × `nitrogen` table is more interesting. There are two s.e.d.'s according to whether the two means to be compared are for the same variety. If they are, then the sub-plots from which the means are calculated will all involve the same set of whole-plots, so any whole-plot variability will cancel out, giving a smaller s.e.d. than for a pair of means involving different varieties.

Finally, notice that this time the Information output category has generated a message noting that block 1 has a large residual compared to the residuals of the other five blocks. In this instance, the message can be taken as confirming the success of the choice of the

blocks: that is, that the yields of the plots in block 1 are consistently higher than those in the other blocks. Large residuals in the `block.wplot.subplot` stratum, however, might indicate possibly aberrant values.

6.9 Practical

In an experiment to study the effect of two meat-tenderizing chemicals, the two (back) legs were taken from four carcasses of beef and one leg was treated with chemical 1 and the other with chemical 2. Three sections were then cut from each leg and allocated (at random) to three cooking temperatures, all 24 sections (4 carcasses × 2 legs × 3 sections) being cooked in separate ovens. The table below shows the force required to break a strip of meat taken from each of the cooked sections (the data are also in the file `Meat.gsh`). Analyse the experiment.

Leg		1			2		
Carcass	Section	Chemical	Temp	Force	Chemical	Temp	Force
1	1	1	2	5.5	2	3	6.3
	2	1	3	6.5	2	1	3.5
	3	1	1	4.3	2	2	4.8
2	1	2	1	3.2	1	3	6.2
	2	2	3	6.0	1	2	5.0
	3	2	2	4.7	1	1	4.0
3	1	2	1	2.6	1	2	4.6
	2	2	2	4.3	1	1	3.8
	3	2	3	5.6	1	3	5.8
4	1	1	3	5.7	2	2	4.1
	2	1	1	3.7	2	3	5.9
	3	1	2	4.9	2	1	2.9

6.10 Commands for analysis of variance

Most of the menus described in this chapter use the `ANOVA` directive, which analyses *generally balanced* designs. These include most of the commonly occurring experimental designs such as randomized blocks, Latin squares, split plots and other orthogonal designs, as well as designs with balanced confounding, like balanced lattices and balanced incomplete blocks. Many partially balanced designs can also be handled, using pseudo factors, so a very wide range of designs can be analysed.

Before using `ANOVA` we first need to define the model that is to be fitted in the analysis. Potentially this has three parts. The `BLOCKSTRUCTURE` directive defines the "underlying structure" of the design or, equivalently, the *error* terms for the analysis; in the simple cases where there is only a single error term this can be omitted. The `TREATMENTSTRUCTURE` directive specifies the treatment (or *systematic*, or *fixed*) terms for the analysis. The other directive, `COVARIATE`, lists the covariates if an analysis of covariance is required. At the start of a job all these model-definition directives have null settings. However, once any one of them has been used, the defined setting remains in force for all subsequent analyses in the same job until it is redefined.

For example, the statements below were generated by the One-way ANOVA (no Blocking) menu to analyse the example in Section 6.1.

```
"One-way ANOVA (no Blocking)."
```

```
BLOCK "No Blocking"
TREATMENTS diet
COVARIATE "No Covariate"
ANOVA [PRINT=aovtable,information,mean; FPROB=yes] weight
```

The BLOCK (or, in full, BLOCKSTRUCTURE) directive is given a null setting to cancel any existing setting; so this indicates that the design is unstructured and has a single error term. Similarly, the COVARIATE statement cancels any covariates that may have been set in an earlier menu. The TREATMENTS (or, in full, TREATMENTSTRUCTURE) directive is used to specify that we have a single term in the analysis, the main effect of diet.

The first parameter of the ANOVA directive specifies the y-variate to be analysed. The PRINT option is set to a list of strings to select the output to be printed. These are similar to the check boxes of the Further Output menu. The most commonly used settings are:

aovtable	analysis-of-variance table,
information	details of large residuals, non-orthogonality and any aliasing in the model,
covariates	estimated coefficients and standard errors of any covariates,
effects	tables of effects,
residuals	tables of residuals,
contrasts	estimated coefficients of polynomial or other contrasts,
means	tables of means,
%cv	coefficient of variation, and
missingvalues	estimated missing values.

By default PRINT=aovtable,information,covariates,means,missing.

Probabilities are not printed by default for the variance ratios in the analysis-of-variance table, but these can be requested by setting the FPROBABILITY option to yes. ANOVA has a PSE option to control the standard errors printed for tables of means. The default setting is differences, which gives standard errors of differences of means. The setting means produces standard errors of means, LSD produces least significant differences and by setting PSE=* the standard errors can be suppressed altogether. The LSDLEVEL option allows the significance level for the least significant differences to be changed from the default of 5%. ANOVA also has a FACTORIAL option which can be used to specify the maximum order (that is, number of factors) in the treatment terms to be fitted in the analysis; the default is 3.

To show a more complicated example, these statements were generated to analyse the split-plot design in Section 6.8

```
"Split-Plot Design."
BLOCK blocks/wplots/subplots
TREATMENTS nitrogen*variety
COVARIATE "No Covariate"
ANOVA [PRINt=aovtable,information,mean; FACT=3; FPROB=yes]\
  yield
```

The block formula

```
blocks/wplots/subplots
```

expands, as explained in Section 6.7, to give the three terms

```
block + block.wplot + block.wplot.subplot
```

each of which defines a stratum for the analysis. Similarly, the treatment formula

```
nitrogen*variety
```

expands to

```
nitrogen + variety + nitrogen.variety
```

to request that Genstat fits the main effects of nitrogen and variety, and their interaction. Again there are no covariates.

The Further Output menu uses the ADISPLAY directive to produce the output, procedure APLOT to produce the plots of residuals, procedure AGRAPH to plot tables of means, and procedure APERMTEST for permutation tests. ADISPLAY has options PRINT, FPROBABILITY, PSE and LSDLEVEL like those of ANOVA. However, with ADISPLAY the default for PRINT is to print nothing.

Finally, the AKEEP directive is used by the ANOVA Save Options menu to save the residuals and fitted values after an analysis. This is done by two options called RESIDUALS and FITTEDVALUES. AKEEP also allows information to be saved for any of the individual terms in the analysis. The terms are defined by a formula which is specified using the TERMS parameter. The formula is expanded into a list of model terms, subject to the limit defined by the FACTORIAL option which operates like the FACTORIAL option of ANOVA; the other parameters then specify data structures in parallel with this list, to store the information required. Tables of means are saved using the MEANS parameter. So, for example, the variate of residuals yieldres and the N by S table of means meantab in Figure 6.9 were saved by the statement

```
AKEEP [RESIDUALS=yieldres] N.S; MEANS=meantab
```

Other useful parameters of AKEEP are EFFECTS (tables of effects for treatment terms), REPLICATIONS (replication tables), RESIDUALS (tables of residuals for block terms), DF (degrees of freedom) and SS (sums of squares).

Below we use AKEEP to save the sum of squares and degrees of freedom for nitrogen and variety from the analysis of the split-plot design in Section 6.8.

```
AKEEP nitrogen+variety; SS=N_ss,V_ss; DF=N_df,V_df
PRINT N_ss,N_df,V_ss,V_df; DECIMALS=1,0
```

N_ss	N_df	V_ss	V_df
20020.5	3	1786.4	2

6.11 Other facilities

In this chapter we have shown only four of the 14 design types (including synonyms) that can be analysed using the Analysis of Variance menu. Other possibilities include Latin squares, Graeco-Latin squares, strip-plot and split-split plot designs and lattices.

There are also menus for generating standard designs for analysis later by the analysis of variance menus, as well as for many more-specialized designs such as factorial designs (with treatment terms confounded with blocks), fractional factorial designs (with blocking), alpha designs, balanced-incomplete-block designs, cyclic designs, Box-Behnken designs, central-composite designs, neighbour-balanced designs, Plackett-

Burman designs, loop designs and reference-level designs. Other menus are available to determine sample sizes for t-tests, sign tests, Mann-Whitney tests, binomial tests, correlation coefficients, Lin's concordance coefficient and McNemar's test, also for tests of equivalence and noninferiority made by analysis of variance or t-test. See Chapter 6 of the *Guide to ANOVA and Design in Genstat*, and Sections 4.8 and 4.11 of the *Guide to the Genstat Command Language, Part 2 Statistics*. There are also many design tools, described in Section 4.12 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

The Analysis of Variance menus deal mainly with balanced designs. This ideal situation, however, is not always achievable. The randomized-block design in Section 6.5 is balanced because every block contained one of each treatment combination. However, there may sometimes be so many treatments that the blocks would become unrealistically large. Designs where each block contains less than the full set of treatments include cyclic designs and Alpha designs (both of which can be generated within Genstat by clicking Stats on the menu bar, selecting Design and then Select Design), neither of which tend to be balanced. In experiments on animals, some subjects may fail to complete the experiment for reasons unconnected with the treatments. So even an initially balanced experiment may not yield a balanced set of data for analysis. These situations are handled by the Mixed Models (REML) menus, which use the Genstat `REML` directive. They also allow you to fit models to the complex correlation structures that occur in repeated measurements or in spatially-correlated data from field experiments. See the *Guide to REML in Genstat*, and Chapter 5 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

7 Other statistical methods

This *Introduction* covers most of the data management and manipulation menus, but gives only a brief introduction to the statistical menus. The additional areas are summarized below, with cross references to other Genstat documentation.

Full details of all the facilities in Genstat, and its commands, are accessible in PDF format from the Help menu (see Figure 1.6). The *Reference Manual* consists of: *1. Summary*, *2. Directives* and *3. Procedures in Procedure Library PL23*. The *Guide to the Genstat Command Language* contains *1. Syntax and Data Management*, and *2. Statistics*. There are also specialized Guides for *ANOVA and Design*, for *Regression, Nonlinear and Generalized Linear Models*, for *REML* (analysis of linear mixed models), for *Multivariate Analysis* and for the *Genstat Spreadsheet*.

7.1 Mixed models (REML)

The REML menus allow you to analyse linear mixed models i.e. linear models that can contain both fixed and random effects. In some applications these are known as "multi-level" models. It can thus be used to analyse unbalanced designs with several error terms (which cannot be analysed by the analysis of variance menus).

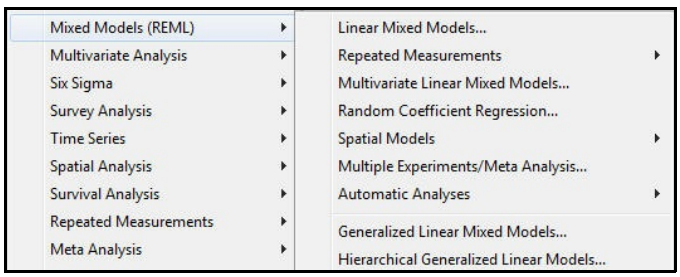


Figure 7.1

REML can also fit random correlation models to describe the covariances between random effects. This allows you to fit spatial correlation models to describe the random variation in large field experiments. You can use the Spatial Model menus to investigate potential models yourself, or the menus for Automatic Analyses to get Genstat to find an appropriate random model for you automatically. For further details see the *Guide to REML in Genstat*.

7.2 Multivariate analysis

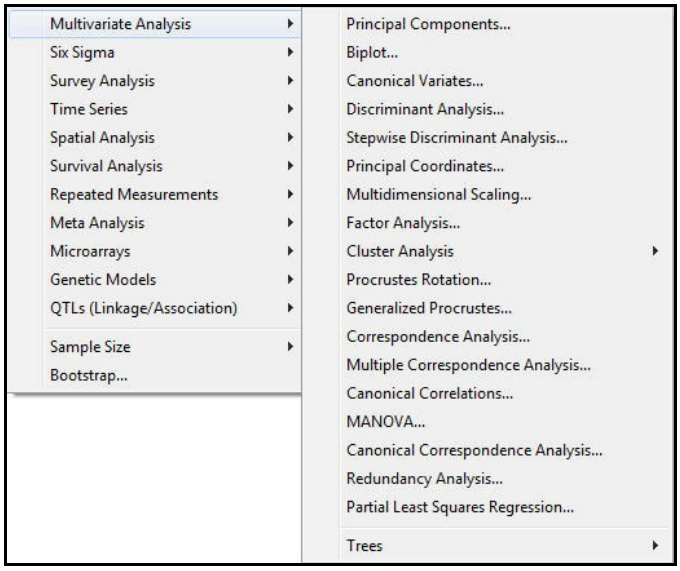


Figure 7.2

Multivariate analysis is useful when you have several different measurements on a set of *n* objects. In Genstat the measurements would usually be stored in separate variates, and these would have a unit for each object. The objects are often regarded as being a set of *n* points in *p* dimensions (*p* being the number of variates).

Many techniques, for example principal components analysis and canonical variates analysis are aimed at reducing the dimensionality. That is, they aim to find a smaller number of dimensions (usually 2 or 3) that exhibit most of the variation present in the data. This can help you determine patterns or structure in the data, as well as identify the relative importance of individual variables. Genstat has several menus for producing graphical representations, for example multidimensional scaling and principal coordinates analysis. It also has facilities for modelling multivariate data, including multivariate analysis of variance and partial least squares.

Another important requirement is to take a set of units and classify them into groups based on their observed characteristics. Hierarchical cluster analysis starts with a set of groups each of which contains one of the units. These initial groups are successively merged into larger groups, according to their similarity, until there is just one group containing all the observations. Genstat also provide menus for non-hierarchical classification, where the aim is to form a single grouping of the observations that optimizes some criterion such as the within-class dispersion, or the Mahalanobis squared distance between the groups, or the between-group sum of squares.

Other facilities include those for constructing classification trees, which allow you to predict the classification of unknown objects using multivariate observations.

For further details see the *Guide to Multivariate Analysis in Genstat* and Chapter 6 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

7.3 Time series

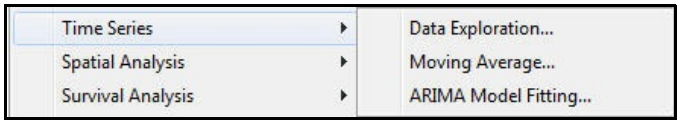


Figure 7.3

A *time series* in Genstat is a variate containing a sequence of observations made at equally spaced points in time. The time series menus (Figure 7.2) are designed to allow you first to explore the data by plotting graphs and printing autocorrelations, and then to fit *autoregressive integrated moving-average* (ARIMA)

models as advocated by Box & Jenkins (1970). Details of the full time series facilities in Genstat, and information about the underlying theory, are given in Chapter 7 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

7.4 Six sigma

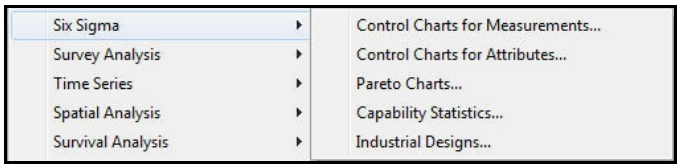


Figure 7.4

Genstat has wide range of facilities to support the six-sigma approach to quality improvement. There are some specialized menus in the Six Sigma option of the Stats menu on the menu bar. These include control charts, Pareto charts and capability statistics, and a wizard to form various designs popular in industrial statistics. Further details are in Section 2.10 of Part 2 of the *Guide to the Genstat Command Language*.

7.5 Survey data

The Survey Analysis option of the Stats menu (Figure 7.4) provides a suite of menus for the analysis of simple or complex surveys, including modelling and imputation. Full details are in the *Guide to Survey Analysis in Genstat*.

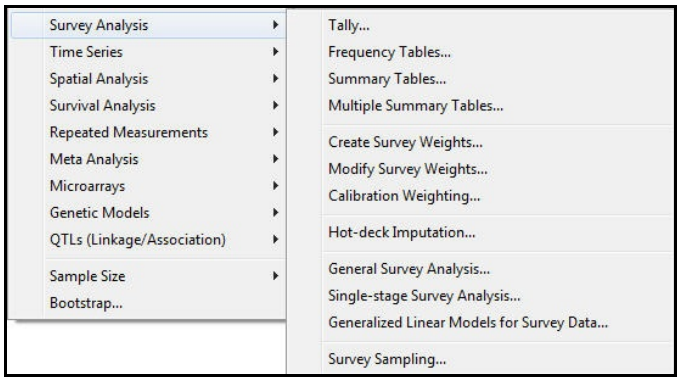


Figure 7.5

7.6 Geostatistics

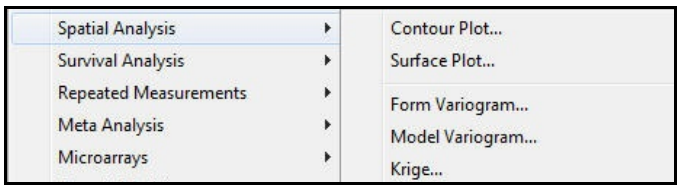


Figure 7.6

Genstat has a set of menus in the Spatial Analysis option of the Stats menu for spatial analysis by "kriging". This is a method originating in geostatistics for analysing data distributed in two dimensions. The kriging model specifies how successive measurements of a variable in space are correlated with each other, in terms of a "variogram". This is analogous to the "correlogram" used in the analysis of time series, but for two-dimensional (spatial) data rather than one-dimensional (temporal) data. Genstat has a menu to form the variogram (see Figure 7.5). There is then a menu for fitting models to describe how the correlations vary with distance, and perhaps also with direction. Finally, the Krige menu allows you to generate predictions (and their variances). Examples, and a description of the underlying methodology, are in Section 8.3 of Part 2 of the *Guide to the Genstat Command Language*.

7.7 Survival analysis

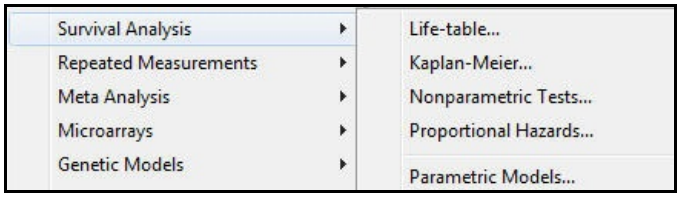


Figure 7.7

Survival data are data in which the response variate is, for example, the lifetime of a component or the survival time of a patient. Typically these are censored, i.e. some individuals survive beyond the end of the study, and so their survival time is unknown. The survivor function F(t) is defined as the probability that an individual is still surviving at time t. The Kaplan-Meier estimate of the survivor function (provided by the Kaplan-Meier menu) is simply the number surviving out of the number at risk in each time interval. Alternatively, Genstat can calculate the life-table (or *actuarial*) estimates of the survivor function. Nonparametric tests can be made to compare the survival distributions of two or more groups of right-censored survival data. There are also menus for modelling the survival times, by assuming that they follow exponential, Weibull or extreme-value distributions, or by fitting proportional hazards models. Further details are in Section 8.2 of Part 2 of the *Guide to the Genstat Command Language*.

7.8 Repeated measurements

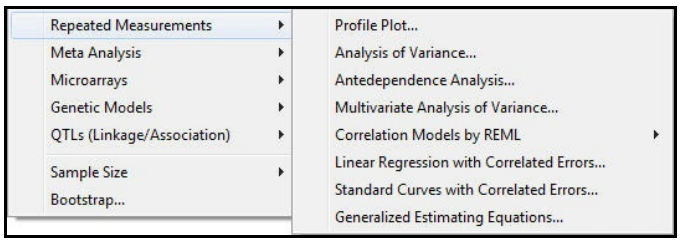


Figure 7.8

A repeated-measurements study is one in which subjects (animals, people, plots, etc) are observed on several occasions. Each subject usually receives some randomly allocated treatment, either at the outset or repeatedly through the investigation, and is then observed at successive occasions to see how the treatment effects develop. Genstat has a comprehensive collection of menus for the analysis of such data (see Figure 7.7). You can model the correlation structure over time using REML; see Chapter 4 of the *Guide to REML in Genstat* or Sections 5.4 of Part 2 of the *Guide to the Genstat Command Language*. Other menus provide customized plotting of the observations (or *profiles*) against time, repeated measures analysis of variance, analyses based on ante-dependence structure or generalized estimating equations, and regression or nonlinear modelling where the residuals follow an AR1 or power-distance correlation model. For details see Sections 8.1 and 3.5.12 of Part 2 of the *Guide to the Genstat Command Language*.

7.9 Meta analysis

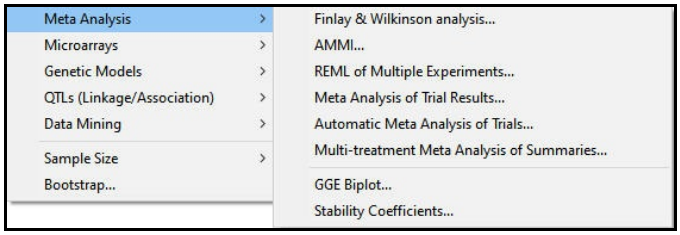


Figure 7.9

The meta analysis menus (Figure 7.8) provide analyses that combine information from several related experiments. This process, often called meta analysis, can be performed using the REML methods and menus; see Chapter 2 of the *Guide to REML in Genstat* or Section 5.7 of Part 2 of the *Guide to the Genstat Command Language*. There are also has menus for fitting AMMI (additive main effects and multiplicative interaction) and Finlay and Wilkinson models, and for the more standard meta analysis where combined estimates of treatment effects are generated from the results of individual trials. Finally, there are two menus that may be useful in the analysis of variety trials. The GGE Biplot menu provides a range of plots that are useful for assessing the performance of genotypes in different environment. Essentially these are standard principal-component biplots, but various additional information can be added to the plots, as suggested in the book *GGE Biplot Analysis* by Yan & Kang (2003), to help elucidate the genotype and environment relationships. The Stability Coefficients menu calculates stability coefficients for genotype-by-environment data. For further details, see Part 3 of the *Genstat Reference Manual* (procedures `AMMI`, `META`, `GGEBIPLOT`, `GESTABILITY` and `RFINLAYWILKINSON`).

7.10 Microarray data

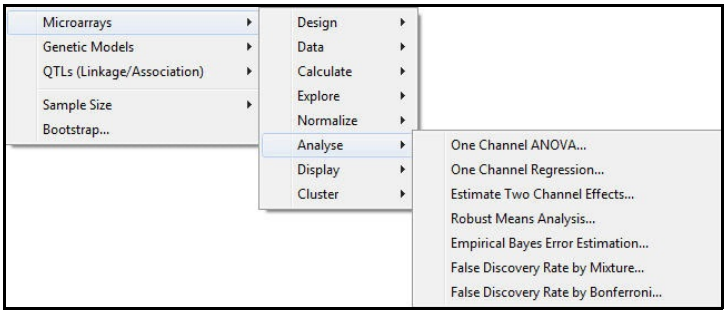


Figure 7.10

The microarrays menus (Figure 7.9) read data in multiple formats (GenePix, Imagene, Spot, TIGR MEV, ScanAlyze, QuantArray, Affymetrix and generic CSV files) and calculate the expression values with a wide range of options for estimation and background correction. The data can be visualized with customized graphics procedures. Affymetrix data can be normalized by quantile normalization. Data from 2-colour microarrays can be normalized with either loess or spline models, where the spline models also allow for autoregressive errors and thin plate splines for the row-by-column effects. The normalized results can then be put through a regression model to estimate the treatment effects from an experiment, with allowance for dye bias. An empirical Bayes model can be used for the distribution of the estimated standard errors over the set of probes, to obtain more robust estimates of the probability of a probe being differentially expressed. Estimates of False Discovery Rates can be made using a Beta-uniform mixture model. The results can then be clustered using any of the standard cluster algorithms available

in Genstat. These menus provide a straightforward road map through all the steps in the analysis of microarray data.

7.11 QTL analysis

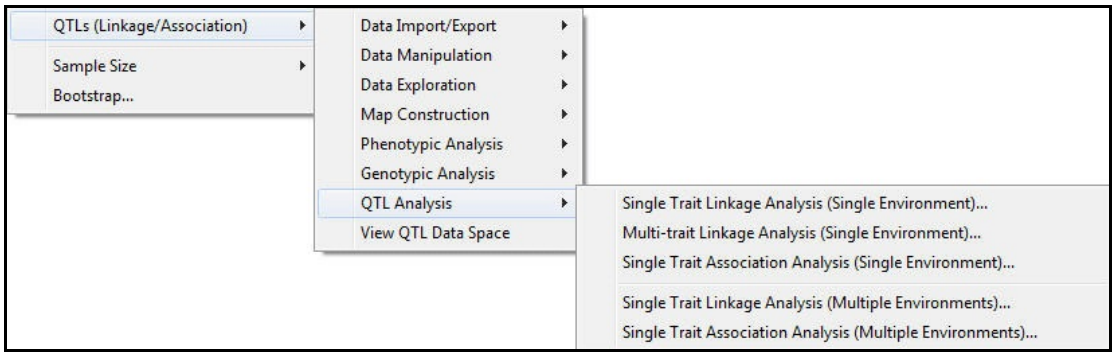


Figure 7.11

The QTL menus provide access to a wide range of analyses and displays to allow the estimation of quantitative trait loci (QTLs) from single environment or multi-environment trials, and for map construction. You can access them from the Stats menu (Figure 7.10).

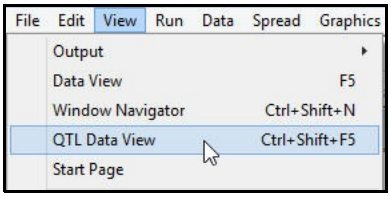


Figure 7.12

Alternatively, if you click on the QTL Data View option of the View menu (Figure 7.11) on the menu bar, a special QTL Data view tab will become available alongside the Data and Window Navigator views; see Figure 7.12. This allows you both to see the available data and to select analyses.

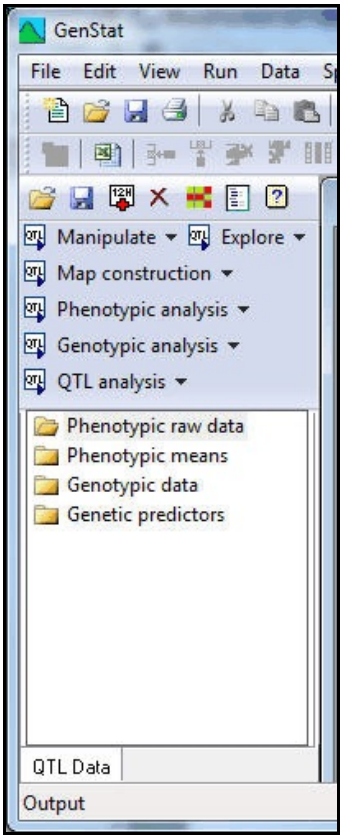


Figure 7.13

7.12 Exact tests

An exact test is a nonparametric test in which the significance levels are calculated without making any assumptions about the probability distributions that generated the observed data values.

For example, consider the two-sample t-test. Under the "null hypothesis" that there is no difference between the two samples, the probability is calculated by assuming that the data values come from Normal distributions with equal means and variances. This can be justified by the observation that, in practice, many samples do seem to come from Normal distributions. A more rigorous justification is based on the fact that the test assesses the difference between the means of the two samples, and means of samples from most distributions tend to become Normally distributed as the sample size becomes increasingly large. Statistically these distributions are said to be asymptotically Normally distributed. So the probability is based on the asymptotic properties of the test (although it usually works well with smaller samples too).

The exact alternative to the conventional t-test makes the assumption that the observed data are representative of the full population of possible data values, and calculates the significance level by considering all the possible ways in which the values could have been allocated to the two samples (including the allocation that actually occurred). The t-statistic is calculated for all of these possibilities, and the probability of the observed data is calculated by seeing where its t-statistic occurs within the full set of statistics. So, for example, it would be significant at the 1% level in a one-sided test if its statistic was in the largest 1% of the statistics.

Genstat can produce exact probabilities for most nonparametric tests, including the Mann-Whitney test, Wilcoxon test, binomial test, sign test, Poisson test, McNemar test, Cochran's Q test and Kendall's t, as well as Fisher's exact test for counts in 2×2 tables (which was the original exact test). For some of the tests, it may not be feasible to calculate the exact probability with very large samples. So there the probability will be based on the asymptotic properties of the test as discussed earlier. However, these are exactly the situations where the asymptotic probabilities can be relied on.

Genstat can also do permutation tests for t-tests, analysis of variance, Steel's test, regression analyses and the analysis of similarities. You specify how many random permutations to make and, if that is greater than the number that is possible for the data set, the exact test is done instead.

Index