GENSTAT DISCOVERY EDITION FOR EVERYDAY USE

Flemish Association for Development Co-Operation and Technical Assistance (VVOB)

World Agroforestry Centre (ICRAF)
Research Support Unit

University of Reading
Statistical Services Centre

GenStat Discovery Edition for everyday use

Wim Buysse, Roger Stern and Ric Coe

Correct citation : Wim Buysse, Roger Stern and Ric Coe, 2004. GenStat Discovery Edition for everyday use . ICRAF Nairobi, Kenya. 114 pp.

ISBN 92 9059 158 7

The publishers and the authors make no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. The publishers and the authors cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

The University of Reading
Whiteknights
PO Box 217
Reading
Berkshire RG6 6AH
United Kingdom
http://www.reading.ac.uk

The World Agroforestry Centre
ICRAF
PO Box 30677
00100 Nairobi

Kenya
http://www.worldagroforestrycentre.org/rsu

Contents

1 GenStat Discovery Edition

1.1 What is this guide about?

This guide is intended primarily for scientists who wish to use GenStat for the analysis of their research data. Most of the examples are taken from the book '*Statistical Methods in Agriculture and Experimental Biology*' by *Mead, Curnow and Hasted*[1], others come from course material developed by ICRAF and the University of Reading. Our primary aim is to teach GenStat, rather than statistics. In some chapter however we review some basic statistics and show how GenStat Discovery Edition can be used to teach statistics. Nevertheless, minimal information is given regarding the data and the interpretation of the results.

Chapter 2 of this guide gives a basic introduction to GenStat and can be considered as a tutorial. It covers data input, some descriptive statistics, calculations and an introduction to the command language. Chapter 3 introduces the application of simple statistical ideas in GenStat (t-test and simple regression). Chapter 6 is about data organization and exploration. Analysis of Variance using GenStat is covered in chapter 8. Examples include a simple randomised block, factorial treatment structure and a split plot design. The other chapters contain review questions or "challenges".

Our main purpose in writing this guide is to provide supporting material for scientists, who are on a training course in statistics. This guide, particularly chapters 1 - 5, may also be used for self-study, either within a supervised environment, or for users who have experience of other statistical packages. This guide is not intended for self-study by beginners to statistical computing.

We estimate that the whole of the guide could be covered in a one-day session on a training course for those familiar with other statistical software. This session would introduce the software and could include a discussion on initial impressions of GenStat at the end of the session. On training courses for participants with limited computer skills, this guide will take at least 4 days. Such training course could include other exercises with additional datasets. Datasets for a course on analysing agroforestry experiments can be found at: http://www.worldagroforestrycentre.org/sites/RSU/dataanalysis/index.asp

All datasets used in the examples and exercises can be found on the CD-rom. If you read this manual in a printed version, the files can be downloaded from the websites of the Research Support Unit of the World Agroforestry Centre (http://www.worldagroforestrycentre.org/rsu). If you read this as a pdf file, you can open the files by clicking on the attachment icon.

[1] Roger Mead, Robert N. Curnow, Anne M. Hasted, 2003. Statistical Methods in Agriculture and Experimental Biology. Third Edition. Chapman & Hall/CRC. 472 pages ISBN 1-58488-187-9

1.2 The origins of the Discovery Edition.

The version of GenStat described here is the GenStat for Windows Discovery Edition. It is based on the Fifth Edition, Service Pack 2 but with the older graphics release 4.1.

The Discovery Edition and this guide are the result of a unique public-private partnership between a software company, research institutes and a development cooperation association. Researchers know that effective statistical analysis is an essential part of their research and needs high quality software. In developing countries, there is often a lack of resources to obtain such high quality software. During a meeting at the GenStat User Conference held in Oxford in September 2001, VSN International Ltd., the distributor of GenStat, was asked to consider the possibility of giving the software for free to researchers in developing countries. At first, VSN International Ltd. was afraid of committing commercial suicide. But gradually ideas changed and on 17 October 2003 GenStat Discovery Edition was officially launched. During a pilot year, GenStat Discovery Edition will be freely available to non-commercial users throughout Africa and will be distributed with extensive online documentation, guides and training material. This is a pilot scheme, which we expect to continue. The edition is supported by the Statistical Services Centre (University of Reading, UK), The World Agroforestry Centre (ICRAF, Kenya), The International Livestock Research Institute (ILRI, Kenya) and the Biometry Unit Consultancy Service (BUCS, University of Nairobi, Kenya). They issue the licenses and develop training materials. To provide assistance to initiatives that improve the situation of available hard and software in the region is one of the objectives of the project "Capacity strengthening in research methods of partners of the World Agroforestry Centre (ICRAF) in East and Central Africa. This project is funded by VVOB, the Flemish Association for Development Cooperation and Technical Assistance. Distributing free high quality software and training materials to non-commercial users through Africa fits very well in this project and VVOB funded the development of a website, part of the development of this guide and part of the distribution of CDs with the software and training materials.

The latest information on the GenStat Discovery Edition offer can be found on: http://www.worldagroforestrycentre.org/GenStatforafrica

1.3 Configuration.

The minimum recommended configuration under Windows 98 is a Pentium PC with 32 Mb RAM. GenStat is developed by the GenStat Committee of the Statistics Department, IACR-Rothamsted, Harpenden, Hertfordshire AL5 2JQ, UK. GenStat is published and distributed by VSN International Ltd, Wilkinson House, Jordan Hill Road, Oxford OX2 8DR, UK (Tel: +44 (0)1865 511245 – Fax: +44 (0)870 1215653 – http://www.vsn-intl.com - E-mail: info@vsn-intl.com). GenStat is a registered trademark of the Lawes Agricultural Trust.

1.4 Acknowledgements

This manual has been adapted and expanded from the first half of the guide called "Using GenStat for Windows, 5th Edition, in Agriculture and Experimental Biology". This was prepared by staff from the SSC, Reading and ICRAF, Nairobi. It was, in turn based on

original notes prepared by Gillian Arnold and Ruth Butler for an MSc course run by the Department of Agricultural Sciences of the University of Bristol. We are very grateful to them and many others who contributed to earlier versions of this guide.

Last but not least we wish to express thanks to the GenStat Team for making high quality statistical software available for free to users who really need it. The continuation of this initiative depends on the provision of feedback by the users.

1.5 Reference for GenStat Discovery Edition

The correct citation when referring to GenStat Discovery Edition in a publication is:

GenStat, 2003. GenStat for Windows. Release 4.23DE Discovery Edition. VSN International Ltd., Hemel Hempstead, UK.

2 GenStat Basics

The aim of this introductory chapter is to become familiar with the basics of how GenStat for Windows operates.

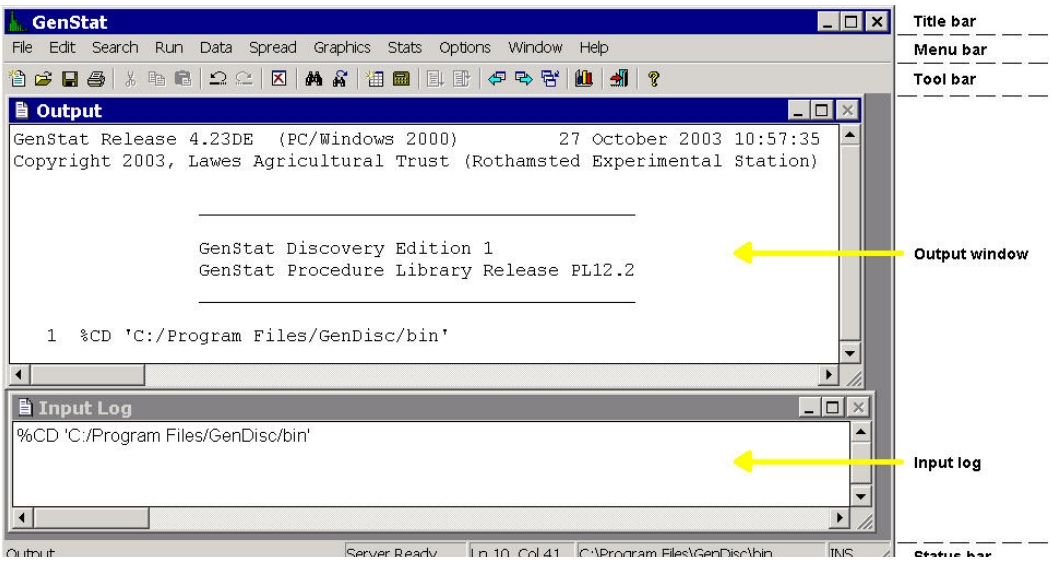
In this manual, we sometimes assume a user who already has experience of MS Excel since most users will have organised their data in a spreadsheet and MS Excel is currently the most widely used spreadsheet. We show how data entered into Excel can be analysed with GenStat and also how data from GenStat can be saved as an Excel file. However, experience with Excel is not necessary for using GenStat.

2.1 Starting GenStat Discovery Edition

Once GenStat Discovery Edition is installed and you have obtained the free license key, you start GenStat Discovery Edition within Windows on a PC by clicking on the GenStat icon on the desktop or toolbar or by selecting the GenStat executable, from the Programs Menu. If no GenStat icon is available on the desktop, you can create one yourself[1].

GenStat
Discovery
Edition

Fig. 2.1 Some of the GenStat windows after start up, also showing several toolbars.



```
GenStat Release 4.23DE  (PC/Windows 2000)         27 October 2003 10:57:35
Copyright 2003, Lawes Agricultural Trust (Rothamsted Experimental Station)

                 _________________________________________

                 GenStat Discovery Edition 1
                 GenStat Procedure Library Release PL12.2
                 _________________________________________

   1  %CD 'C:/Program Files/GenDisc/bin'
```

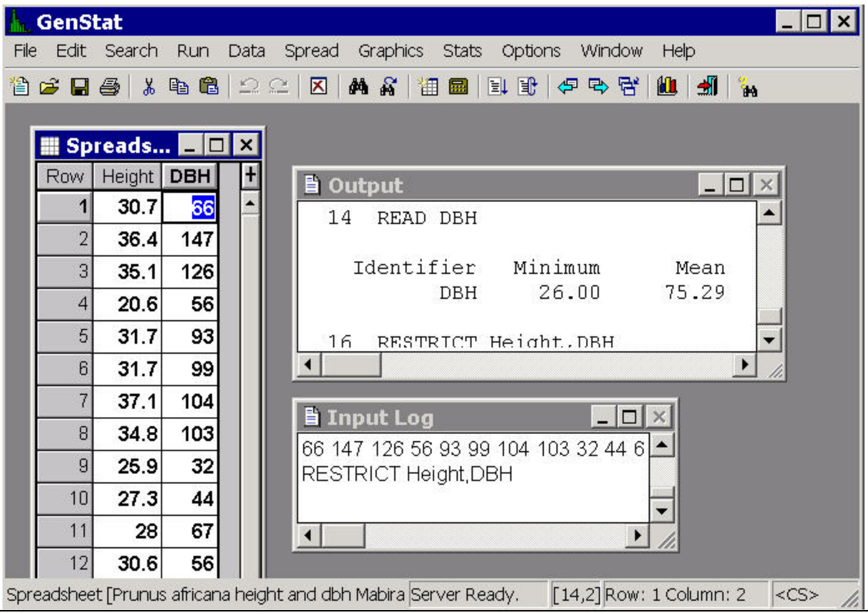
```
%CD 'C:/Program Files/GenDisc/bin'
```

[1] By default, GenStat Discovery Edition is installed in the folder C:\Program files\GenDisc. Use Windows Explorer and go to the subfolder C:\Program files\GenDisc\bin. Right click with the mouse on Genwin42.exe and create a shortcut. This shortcut can now be dragged onto the desktop. You might rename the icon on the desktop (right click on the icon and click 'rename') as GenStat Discovery Edition, to avoid confusion with previous or newer versions.

After starting GenStat, you see a standard Windows interface (Fig. 2.1) with a title bar, menu bar, tool bar, status bar and several windows (Fig. 2.2). The Output window will contain the output from the operations we perform. The input log keeps a record of what has been done in an analysis. Many of the menus are standard for Windows applications. Only Run, Data, Spread, Graphics and Stats are GenStat-specific.

Below an example is given of the GenStat for Windows interface after a spreadsheet has been opened.

Fig. 2.2 Some GenStat windows once data have been entered.



Row	Height	DBH
1	30.7	66
2	36.4	147
3	35.1	126
4	20.6	56
5	31.7	93
6	31.7	99
7	37.1	104
8	34.8	103
9	25.9	32
10	27.3	44
11	28	67
12	30.6	56

```
 14   READ DBH

    Identifier   Minimum       Mean
           DBH     26.00      75.29

 16   RESTRICT Height,DBH
```

```
66 147 126 56 93 99 104 103 32 44 6
RESTRICT Height,DBH
```

Spreadsheet [Prunus africana height and dbh Mabira Server Ready. [14,2] Row: 1 Column: 2 <CS>

2.2 Data input

2.2.1 Data input using the Spread Menu.

We show two ways of entering data into GenStat. The first is with GenStat. Choose ***Spread*** ⇒ ***New*** ⇒ ***Blank*** as in Fig. 2.3.

Fig. 2.3 Spread => New => Blank

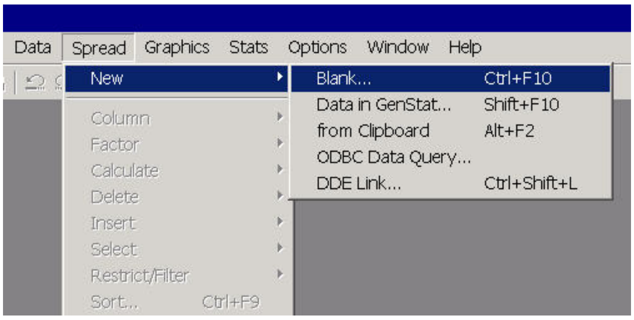
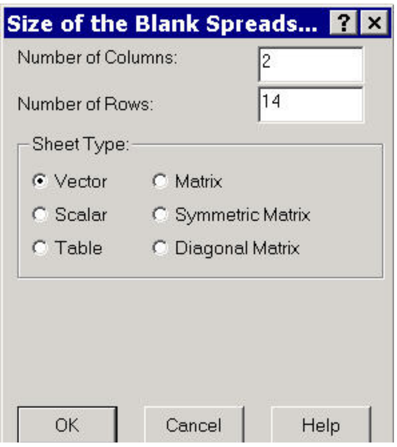


Fig. 2.4 Initial size of the spreadsheet



Number of Columns: 2

Number of Rows: 14

Sheet Type:

● Vector ○ Matrix

○ Scalar ○ Symmetric Matrix

○ Table ○ Diagonal Matrix

Choosing ***Blank*** brings up a box allowing you to specify how many data columns you want, and how many rows of data there will be. Edit the box to make a GenStat spreadsheet with 2 columns and 14 rows as shown in (Fig. 2.4).

Different types of spreadsheet can be made, but the default (i.e. what GenStat will select in the absence of any further information) - ***Vector*** - is usually the type you will need. Click ***[OK]***, and an empty spreadsheet will appear. You can start to enter data by clicking in a cell in the spreadsheet. Type the number, and then press the ***[Enter]*** key. Enter the following numbers into the first column:

30.7 36.4 35.1 20.6 31.7 31.7 37.1 34.8 25.9 27.3 28 30.6 22.3 14.4

Press the ***[Enter]*** key after the last number. The cursor will then move to the top of the next column. Enter these numbers into the second column:

66 147 126 56 93 99 104 103 32 44 67 56 35 26

> Make sure that you press the ***[Enter]*** key after typing the final number, otherwise the content of the last cell will not be send to the GenStat server.

If you have made any mistakes, these can be easily corrected, using the arrow keys to move to the cell to amend and entering the correct value.

For each row, the value in the first column is the height of *Prunus africana* trees in a forest in Uganda. The data were measured as part as a research project of ICRAF. The value in the second column gives the diameter of the same tree. So the first tree is 30.7 metres high and has a diameter at breast height of 66 centimetres.

It is considered a good data management practice to give a **detailed description** of your data. If you save your spreadsheet leaving the column names C1 and C2, some time in the future you will not remember anymore what these data were about. Also none of your colleagues is likely to figure out the meaning of your data.

Adding a detailed description in GenStat can be done in several ways:

- giving a meaningful column name

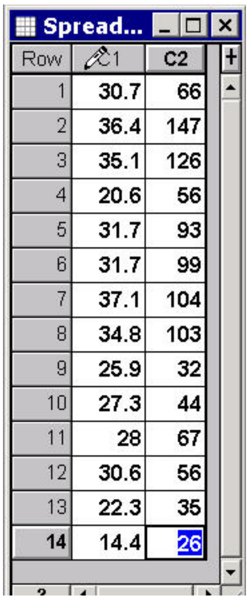
- adding extra description to the column

- giving a meaningful name to the spreadsheet.

2.2.1.1 Naming colums.

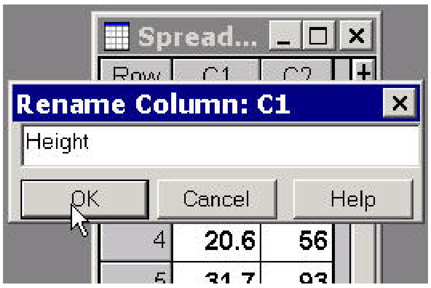
To change the column names from the default C1 and C2 to something more meaningful, position the cursor as shown in the figure Fig. 2.5 below. It becomes a pencil, rather than a hand, and clicking on the mouse gives a popup screen where you can type the name for the column, as shown in Fig. 2.6. Then press ***[OK]***.

Fig. 2.5 Step one in renaming a column



Row	C1	C2
1	30.7	66
2	36.4	147
3	35.1	126
4	20.6	56
5	31.7	93
6	31.7	99
7	37.1	104
8	34.8	103
9	25.9	32
10	27.3	44
11	28	67
12	30.6	56
13	22.3	35
14	14.4	26

Fig. 2.6 Giving the column a new name.

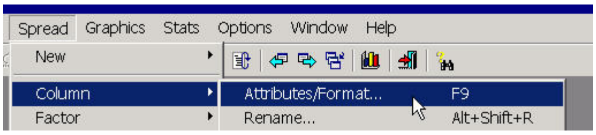
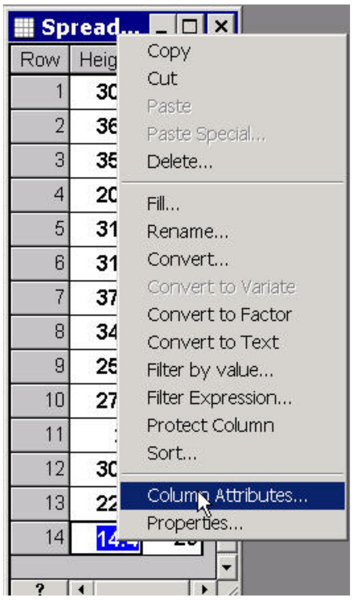


Height

Once you have given column C1 the name “Height”, repeat with C2 with the name “DBH” (for “Diameter at Breast Height). These names now appear on the columns of the spreadsheet.

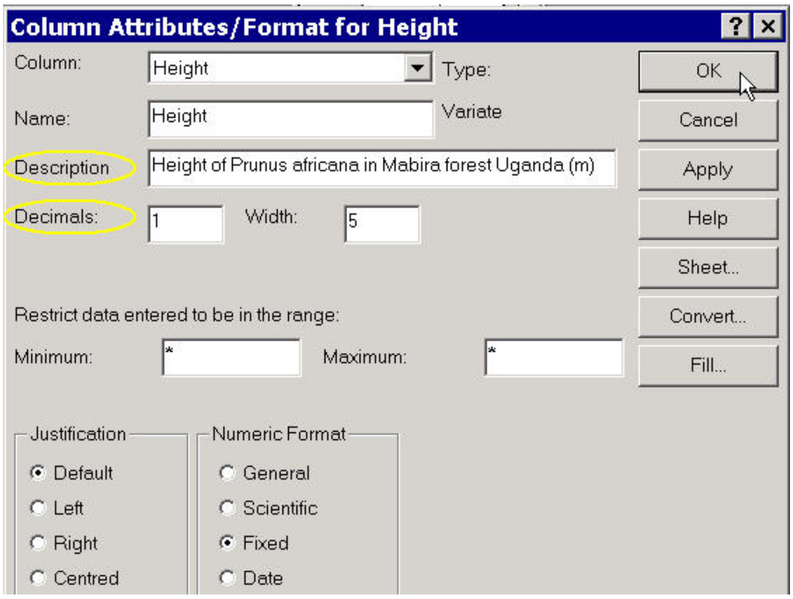
2.2.1.2 Adding extra description.

Another way of changing the column name is to choose ***Spread => Column => Attributes/Format***, see Fig. 2.7, or click in a column and press ***[F9],*** or right-click in a column and choose ***Column attributes*** (Fig. 2.8).

Fig. 2.7 Spread => Column => Attributes/Format	*Fig. 2.8 Right-click for common features*
 Spread Graphics Stats Options Window Help New Column Factor Attributes/Format... F9 Rename... Alt+Shift+R	 Spread... Row Heig Copy Cut Paste Paste Special... Delete... Fill... Rename... Convert... Convert to Variate Convert to Factor Convert to Text Filter by value... Filter Expression... Protect Column Sort... Column Attributes... Properties...

The result in all cases will be a window that gives all kind of information on column attributes and the way the column is formatted (Fig. 2.9). Since some of the data have one decimal, we could enter 1 in the ***Decimals*** box. You can change the name of the column but you can especially add extra information in the ***Description*** box. The description can be maximum 39 characters long.

Fig. 2.9 Formatting the column attributes



Column Attributes/Format for Height

Column: Height

Type: Variate

Name: Height

Description: Height of Prunus africana in Mabira forest Uganda (m)

Decimals: 1

Width: 5

Restrict data entered to be in the range:

Minimum: *

Maximum: *

Justification: ● Default ○ Left ○ Right ○ Centred

Numeric Format: ○ General ○ Scientific ● Fixed ○ Date

This example also shows a general point when using GenStat, namely that there is usually more than one way to call a dialogue. We often find that the quickest route is to right-click, but this only gives the most common features.

2.2.1.3 Naming spreadsheets.

Use ***File => Save As*** to save the file. By default you are prompted to save the file as sheet1.gsh in the C:\Program Files\GenDisc\bin folder (Fig. 2.10). Will you one year from now still be able to remember what is the content of sheet1.gsh? Or will you be able to differentiate it from sheet 453.gsh?

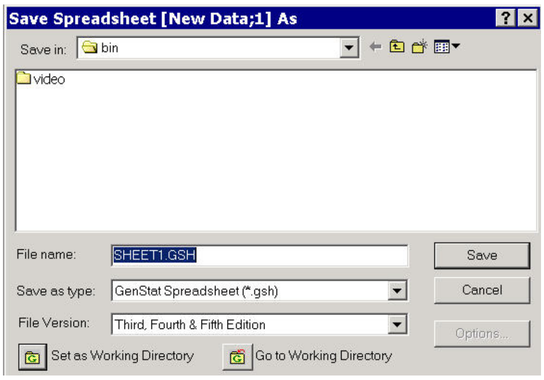
It is a good practice to choose something meaningful as a file name, for instance “Prunus africana height and dbh Mabira Uganda.gsh”. The filename can be anything that is acceptable to your computer system. The Windows 2000 Help gives for instance:

> A file name can contain up to 215 characters, including spaces. However, it is not recommended that you create file names with 215 characters. Most programs cannot interpret extremely long file names. File names cannot contain the following characters: \ / : * ? " < > |

So, use long and descriptive names but don’t exaggerate.

It is recommended to change the working directory (Fig. 2.11). This is the default directory where GenStat will save spreadsheets and other file types. The C:/Program Files/GenDisc/bin directory is used for executables and drivers, so it is better not to mess around in this folder. If you have a D-drive, use Windows explorer to create a folder on that drive for your data files. You can even create one folder per project, each containing several subfolders.

Fig. 2.10 Default directory and name



Save Spreadsheet [New Data;1] As

Save in: bin

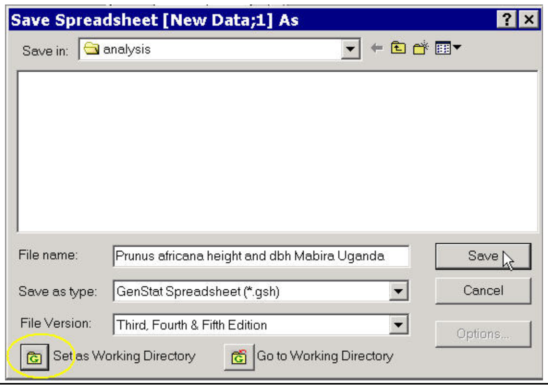
video

File name: SHEET1.GSH

Save as type: GenStat Spreadsheet (*.gsh)

File Version: Third, Fourth & Fifth Edition

Fig. 2.11 Saving a GenStat spreadsheet with an informative name in a directory of your choice



Save Spreadsheet [New Data;1] As

Save in: analysis

File name: Prunus africana height and dbh Mabira Uganda

Save as type: GenStat Spreadsheet (*.gsh)

File Version: Third, Fourth & Fifth Edition

Now use ***Run => Restart Session***, to clear everything from the memory. To continue with the next chapter, minimize GenStat and open MS Excel.

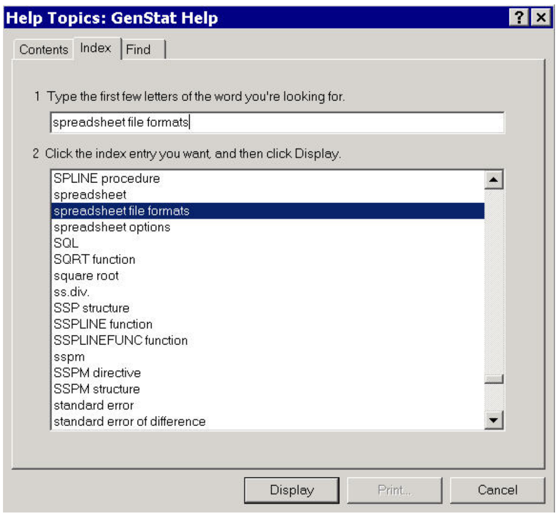
2.2.2 Data input from Excel worksheets.

You may already have your data in a spreadsheet like MS Excel. Importing data from an MS Excel spreadsheet into GenStat is very easy.

> If you entered your data using another software, you still can browse through this section since most procedures will be quite similar. If you are not familiar with MS Excel, you can skip this section.

GenStat can import data from many spreadsheet formats. To know which ones, choose ***Help => Contents and Index*** and type “*spreadsheet*” in the ***Index*** box (Fig. 2.12), select the “*spreadsheet file formats*” entry and click ***[OK]***. The resulting help-file shows information on all possible file formats as in Fig. 2.13.

Fig. 2.12 Help => Contents and Index



Help Topics: GenStat Help

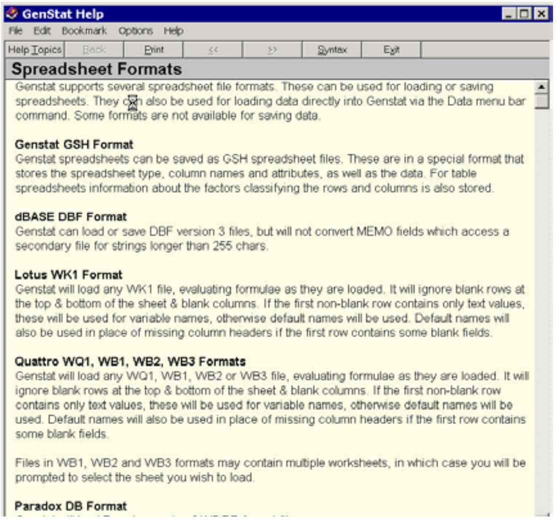
1 Type the first few letters of the word you're looking for.

spreadsheet file formats

2 Click the index entry you want, and then click Display.

SPLINE procedure
spreadsheet
spreadsheet file formats
spreadsheet options
SQL
SQRT function
square root
ss.div.
SSP structure
SSPLINE function
SSPLINEFUNC function
sspm
SSPM directive
SSPM structure
standard error
standard error of difference

Fig. 2.13 Importing from different spreadsheets



Spreadsheet Formats

Genstat supports several spreadsheet file formats. These can be used for loading or saving spreadsheets. They can also be used for loading data directly into Genstat via the Data menu bar command. Some formats are not available for saving data.

Genstat GSH Format

Genstat spreadsheets can be saved as GSH spreadsheet files. These are in a special format that stores the spreadsheet type, column names and attributes, as well as the data. For table spreadsheets information about the factors classifying the rows and columns is also stored.

dBASE DBF Format

Genstat can load or save DBF version 3 files, but will not convert MEMO fields which access a secondary file for strings longer than 255 chars.

Lotus WK1 Format

Genstat will load any WK1 file, evaluating formulae as they are loaded. It will ignore blank rows at the top & bottom of the sheet & blank columns. If the first non-blank row contains only text values, these will be used for variable names, otherwise default names will be used. Default names will also be used in place of missing column headers if the first row contains some blank fields.

Quattro WQ1, WB1, WB2, WB3 Formats

Genstat will load any WQ1, WB1, WB2 or WB3 file, evaluating formulae as they are loaded. It will ignore blank rows at the top & bottom of the sheet & blank columns. If the first non-blank row contains only text values, these will be used for variable names, otherwise default names will be used. Default names will also be used in place of missing column headers if the first row contains some blank fields.

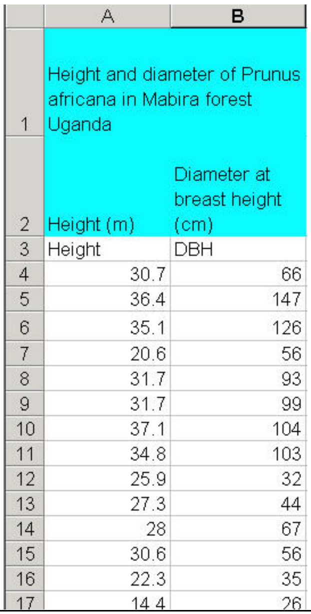
Files in WB1, WB2 and WB3 formats may contain multiple worksheets, in which case you will be prompted to select the sheet you wish to load.

Paradox DB Format

We assume you are now in Excel. Create a new Excel workbook and enter the data from Fig. 2.5. In a spreadsheet like Excel you can add extra information in the cells above the data as in Fig. 2.14:

- You can enter a short name for the column.
- In the row above, you can enter a long name and mention the measurement units.
- One row higher you can add extra information on the experiment.

Fig. 2.14 Data and descriptive information entered into MS Excel.

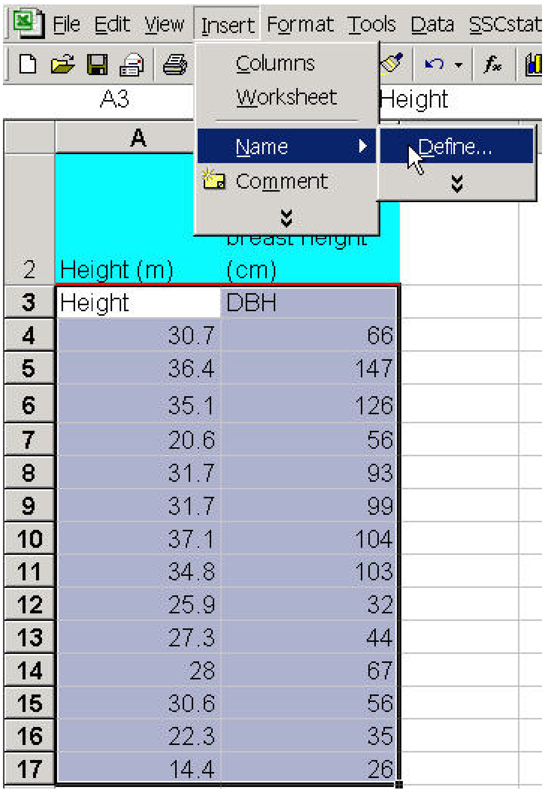


	A	B
1	Height and diameter of Prunus africana in Mabira forest Uganda	
2	Height (m)	Diameter at breast height (cm)
3	Height	DBH
4	30.7	66
5	36.4	147
6	35.1	126
7	20.6	56
8	31.7	93
9	31.7	99
10	37.1	104
11	34.8	103
12	25.9	32
13	27.3	44
14	28	67
15	30.6	56
16	22.3	35
17	14.4	26

The extra information is sometimes called “meta-data” and makes it clear what the data are about. To import the data into GenStat, you define a named range in Excel. In Excel, highlight the range containing the data and the header row and choose ***Insert***

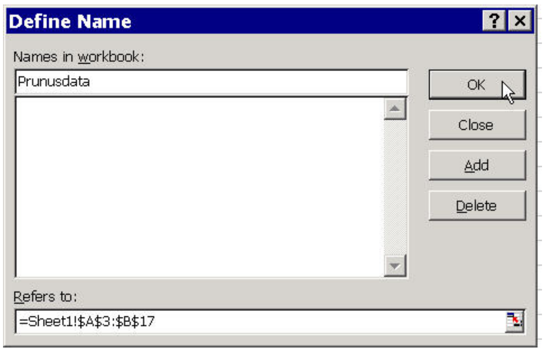
=> Name => Define (Fig. 2.15). Give the range a name, for instance *Prunusdata* (Fig. 2.16). Then save the Excel workbook and give it a meaningful name, for instance "Prunus africana height and dbh Mabira Uganda.xls". You can also rename the worksheet. Right-click on the tab "Sheet 1" and rename it as "Prunus africana" (Fig. 2.17). Then save the Excel file again. You have now finished with Excel, so minimize or close Excel and go back to GenStat.

Fig. 2.15 Defining a named range in Excel



	A	B
2	Height (m)	breast height (cm)
3	Height	DBH
4	30.7	66
5	36.4	147
6	35.1	126
7	20.6	56
8	31.7	93
9	31.7	99
10	37.1	104
11	34.8	103
12	25.9	32
13	27.3	44
14	28	67
15	30.6	56
16	22.3	35
17	14.4	26

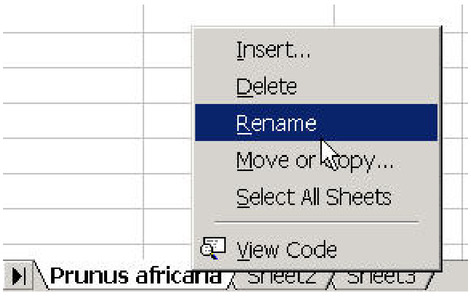
Fig. 2.16 Giving the range a sensible name



Names in workbook: Prunusdata

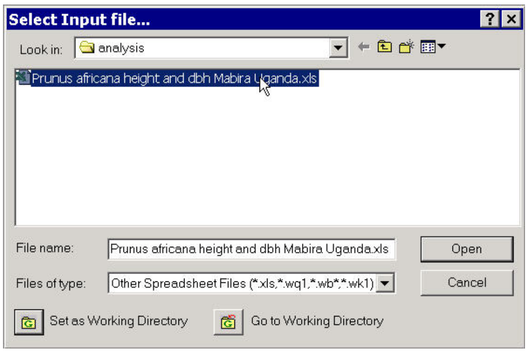
Refers to: =Sheet1!$A$3:$B$17

Fig. 2.17 Naming an Excel sheet



In GenStat, choose ***File => Open*** and select the Input file (Fig. 2.18). You can indicate that the file to import is of the '***Other Spreadsheet Files***' type.

Fig. 2.18 GenStat's File => Open, choosing spreadsheet file-types



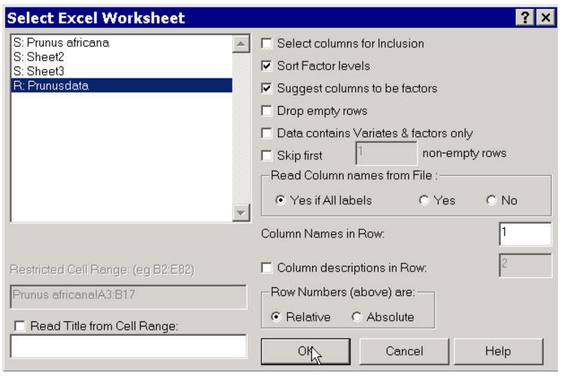
Look in: analysis

Prunus africana height and dbh Mabira Uganda.xls

File name: Prunus africana height and dbh Mabira Uganda.xls

Files of type: Other Spreadsheet Files (*.xls,*.wq1,*.wb*,*.wk1)

Fig. 2.19 Choosing to open the range that was named in Excel, see Fig. 2.16



S: Prunus africana
S: Sheet2
S: Sheet3
R: Prunusdata

□ Select columns for Inclusion
☑ Sort Factor levels
☑ Suggest columns to be factors
□ Drop empty rows
□ Data contains Variates & factors only
□ Skip first 1 non-empty rows

Read Column names from File : ● Yes if All labels ○ Yes ○ No

Column Names in Row: 1

Restricted Cell Range: (eg B2:E82) Prunus africana!A3:B17

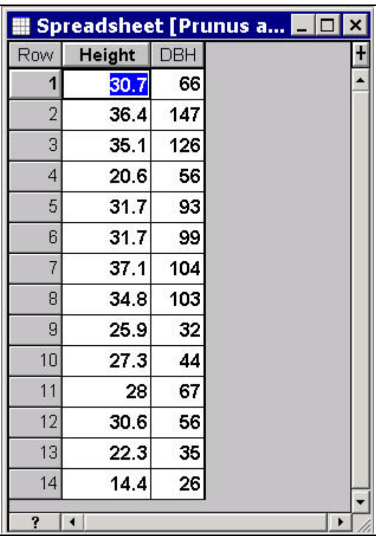
□ Column descriptions in Row: 2

Row Numbers (above) are: ● Relative ○ Absolute

□ Read Title from Cell Range:

In the next window, shown in Fig. 2.19, you can select the named range "*Prunussadata*". The right-hand side of the same window gives various options to customize the way data are imported. By clicking ***[OK]***, the data are immediately imported in a GenStat spreadsheet as shown in Fig. 2.20.

Fig. 2.20 Data imported into a GenStat spreadsheet

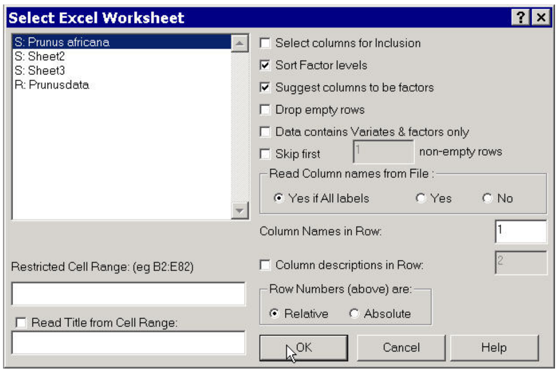


Spreadsheet [Prunus a...

Row	Height	DBH
1	30.7	66
2	36.4	147
3	35.1	126
4	20.6	56
5	31.7	93
6	31.7	99
7	37.1	104
8	34.8	103
9	25.9	32
10	27.3	44
11	28	67
12	30.6	56
13	22.3	35
14	14.4	26

Sometimes people mistakenly import the whole Excel worksheet instead of just the named range. The result is a GenStat spreadsheet that cannot be used, as shown in the following example. In GenStat, restart the session by selecting ***Run ⇒ Restart Session*** and then clicking the ***[Yes]*** button, to clear all windows, dialogue boxes and the spreadsheet. Choose again ***File => Open*** and select the Input file “Prunus africana height and dbh Mabira Uganda.xls”. This time however, select the worksheet “Prunus africana” (Fig. 2.21). The result is a GenStat spreadsheet with 2 columns containing text, as shown in Fig. 2.22. By default, GenStat reads the contents of the Excel cells on the first row as column headers. Since the cells on the second row contain text, GenStat assumes that the whole column contains text and shows a green T next to the column header.

Fig. 2.21 Mistakenly importing a whole sheet from Excel into GenStat



Select Excel Worksheet

S: Prunus africana
S: Sheet2
S: Sheet3
R: Prunusdata

☐ Select columns for Inclusion
☑ Sort Factor levels
☑ Suggest columns to be factors
☐ Drop empty rows
☐ Data contains Variates & factors only
☐ Skip first 1 non-empty rows

Read Column names from File :
● Yes if All labels ○ Yes ○ No

Column Names in Row: 1

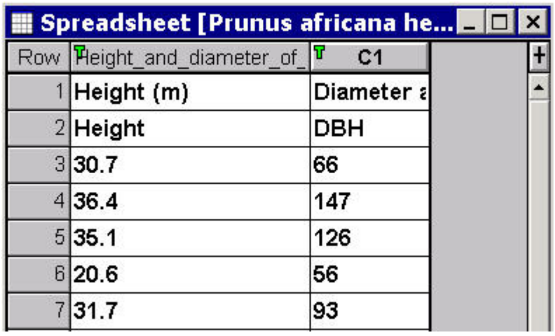
Restricted Cell Range: (eg B2:E82)

☐ Column descriptions in Row: 2

Row Numbers (above) are:
● Relative ○ Absolute

☐ Read Title from Cell Range:

Fig. 2.22 The resulting import includes the column names as data



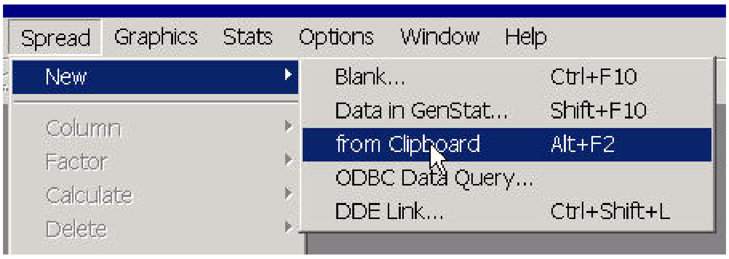
Spreadsheet [Prunus africana he...

Row	T Height_and_diameter_of_	T C1
1	Height (m)	Diameter a
2	Height	DBH
3	30.7	66
4	36.4	147
5	35.1	126
6	20.6	56
7	31.7	93

An alternative way of importing spreadsheet data into GenStat is to copy a range of cells from Excel and paste it into GenStat. This is not good practice in data management, but is a fast and easy way of doing quick provisional analyses

Choose ***Run => Restart Session*** to clear all data out of GenStat. Go back into Excel. Highlight the range containing the data and column headers, right click with the mouse in this range and click ***Copy*** or choose ***Edit => Copy***. Now the data are loaded into the Windows clipboard. Go back to GenStat and choose ***Spread => New => from Clipboard*** (Fig. 2.23) and the data are entered into a GenStat spreadsheet.

Fig. 2.23 Copying data into GenStat from the clipboard



2.2.3 Advanced data input.

If you are going to transfer data repeatedly from the same external file, it is also possible to create links to that file. More information can be found for instance in ICRAF Research Support Unit Technical Note 2, available at

http://www.worldagroforestrycentre.org/sites/RSU/datamanagement/Documents/dupeofduplication.pdf

2.2.4 Leaving GenStat

To end a GenStat session, choose ***File ⇒ Exit***. You will be asked if you want to save any of the open windows or spreadsheets. Select ***[No]*** on all windows and ***[Exit]*** GenStat. More on saving data in different file formats can be found in chapter 2.3.4.

As well as showing you how to enter data into GenStat, you have seen how easy it is to transfer data from another package, such as Excel. So, if you are already familiar with a spreadsheet or another statistical package, using GenStat does not has to stop you from using other software. You can use GenStat in addition. We will show examples from Excel spreadsheets at various points in this guide.

2.3 Some basic data manipulation.

2.3.1 Summary statistics

Restart the session and reopen the file "*Prunus africana height and dbh Mabira Uganda.xls*". The data in the spreadsheet are passed into the GenStat server as soon as you click anywhere outside the spreadsheet or the spread menu. Try doing this by clicking in the output window.

Some summary information about the two columns *Height* and *DBH,* will appear in the output window showing minimum, mean and maximum values, number of values and number of those that are missing.

For further statistical summaries use the ***Stats*** menu, as shown in Fig. 2.24. Choose ***Stats ⇒ Summary Statistics ⇒ Summarise Contents of Variates***. Select the variates to be summarised, as shown in Fig. 2.25, and then click ***[OK]***.

Fig. 2.24 GenStat's descriptive statistics menu

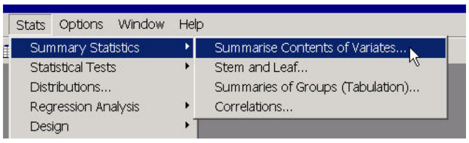
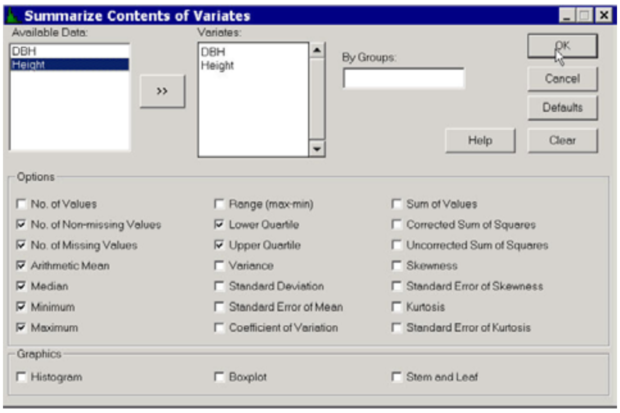


Fig. 2.25 The dialogue to display summary statistics



Available Data: DBH, Height

Variates: DBH, Height

By Groups:

Options

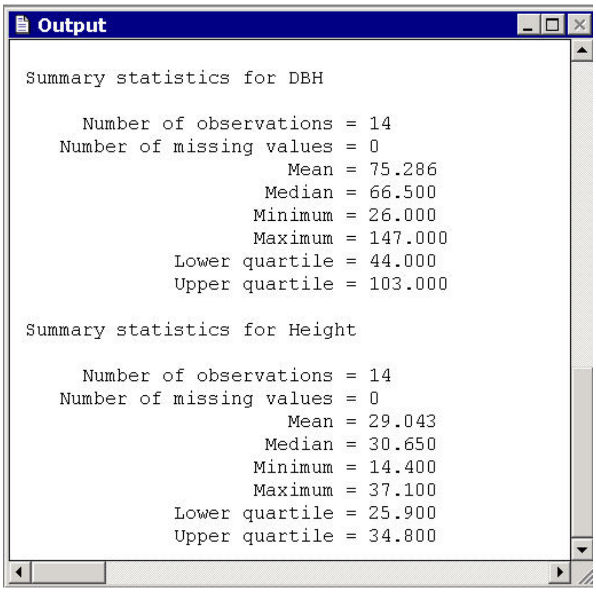
☐ No. of Values
☑ No. of Non-missing Values
☑ No. of Missing Values
☑ Arithmetic Mean
☑ Median
☑ Minimum
☑ Maximum
☐ Range (max-min)
☑ Lower Quartile
☑ Upper Quartile
☐ Variance
☐ Standard Deviation
☐ Standard Error of Mean
☐ Coefficient of Variation
☐ Sum of Values
☐ Corrected Sum of Squares
☐ Uncorrected Sum of Squares
☐ Skewness
☐ Standard Error of Skewness
☐ Kurtosis
☐ Standard Error of Kurtosis

Graphics

☐ Histogram
☐ Boxplot
☐ Stem and Leaf

Select the Output Window. If you cannot see this window, try clicking the or buttons in the toolbar successively until it appears. Some of the results are shown in Fig. 2.26 below.

Fig. 2.26 The default summary statistics



```
Summary statistics for DBH

     Number of observations = 14
   Number of missing values = 0
                       Mean = 75.286
                     Median = 66.500
                    Minimum = 26.000
                    Maximum = 147.000
             Lower quartile = 44.000
             Upper quartile = 103.000

Summary statistics for Height

     Number of observations = 14
   Number of missing values = 0
                       Mean = 29.043
                     Median = 30.650
                    Minimum = 14.400
                    Maximum = 37.100
             Lower quartile = 25.900
             Upper quartile = 34.800
```

There are other statistics available with the dialogue box shown in Fig. 2.25. Find the dialogue box again. Click on the ***[Clear]*** button to clear all currently selected statistics. Reselect the variables and choose Arithmetic Mean, Standard Deviation and Standard Error of Mean, and click ***[OK]***.

In the box in Fig. 2.25, you could already have selected to display a histogram, boxplot and stem and leaf diagram. A range of other graphs is possible with the Graphics menu. Let's see for instance if there is a relationship between height and diameter. Use ***Graphics ⇒ Point Plot*** (see Fig. 2.27) and complete the dialogue box as shown in Fig. 2.28. This will give the scatterplot from (Fig. 2.29).

Fig. 2.27 GenStat's graphics menu

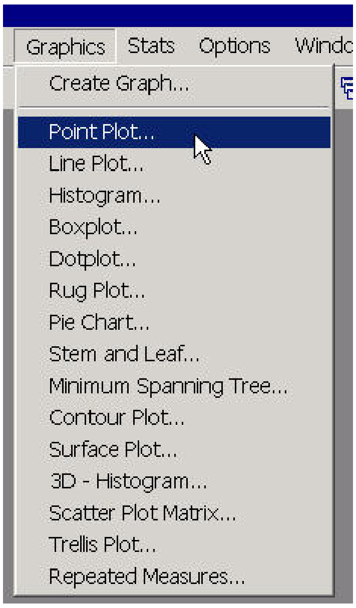
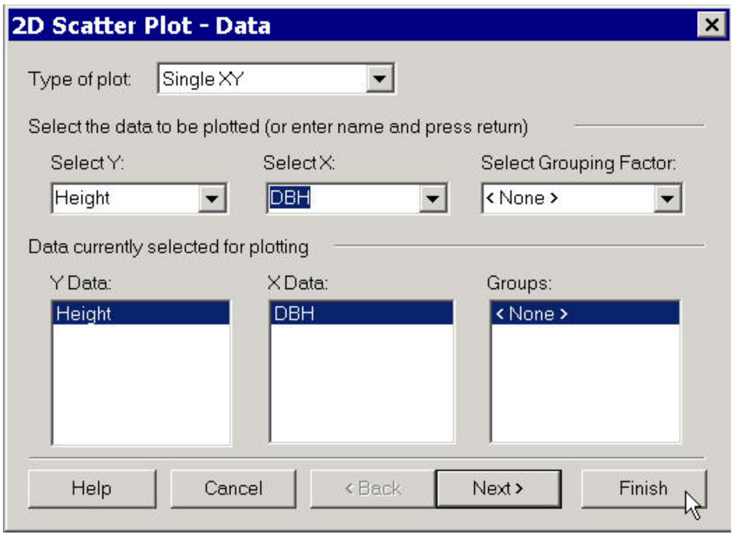


Fig. 2.28 An x-y plot dialogue



2D Scatter Plot - Data

Type of plot: Single XY

Select the data to be plotted (or enter name and press return)

Select Y: Height

Select X: DBH

Select Grouping Factor: < None >

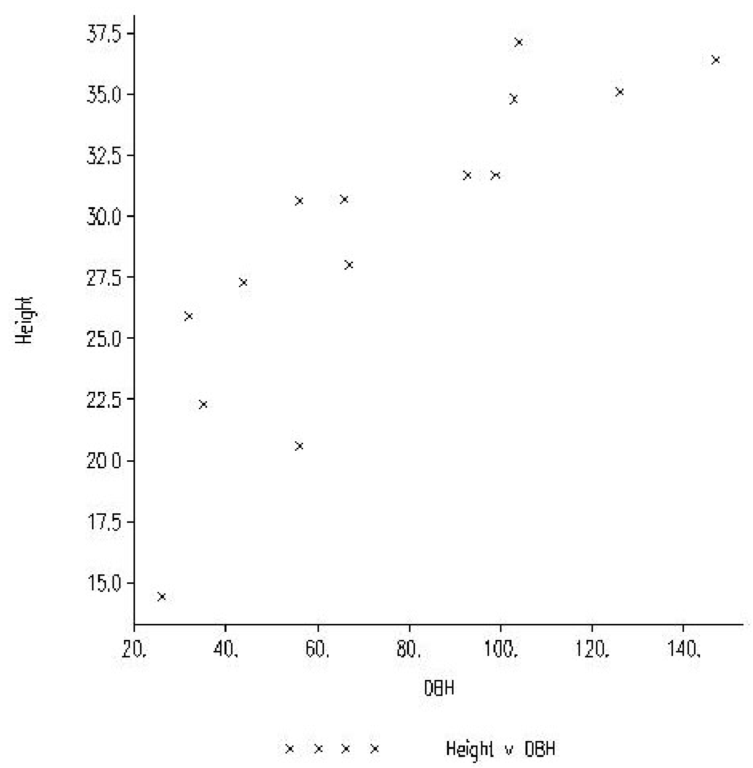
Data currently selected for plotting

Y Data: Height

X Data: DBH

Groups: < None >

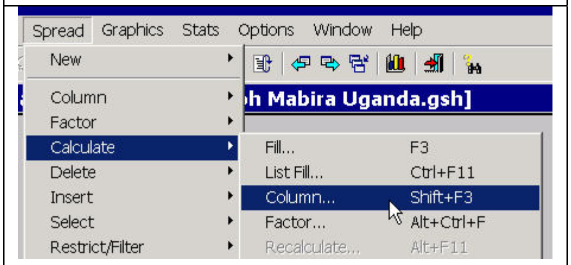
Fig. 2.29 Resulting scatterplot in a separate window



2.3.2 Calculating and formatting columns.

It is easy to calculate new variates from those already entered in a GenStat session. Choose ***Spread => Calculate => Column*** (Fig. 2.30) and select the calculation you need and the name of the new variable that you wish to save.

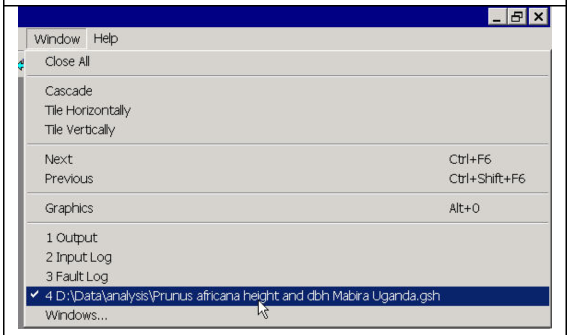
Fig. 2.30 Menu for the GenStat calculator



The next example is not the simplest, but illustrates the ease with which calculations can be done. Often when measuring trees, you want to calculate the volume. The volume of a quadratic paraboloid is often used as an approximation to the volume of the tree. The general formula for this is: ***V= 0.5*g*h*** with g being the basal area and h the height of the tree.

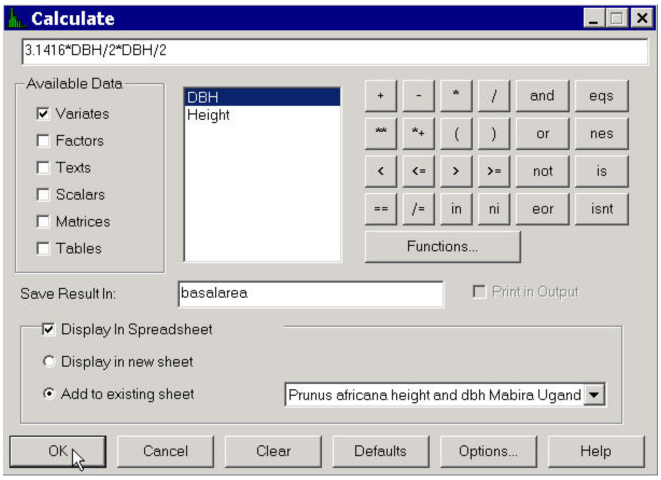
First select the spreadsheet "*Prunus africana height and dbh Mabira Uganda.xls*" again. Do this, either by clicking somewhere in it (if you can see it), or use the toolbar arrow buttons or the ***Window*** menu (Fig. 2.31).

Fig. 2.31 One way of retrieving the spreadsheet



To calculate a new column, choose ***Spread*** ⇒ ***Calculate*** ⇒ ***Column*** as shown in Fig. 2.30. First we calculate a column with the basal area, given by the formula: ***3.1416 * DBH/2 * DBH/2***. Position the cursor in the large box on top of the calculate window before you start typing. You can either type the names of the variables or double click on them in the list with the available data. Also type the name of the new column into the bottom box labelled ***Save Result In*** as shown in Fig. 2.32.

Fig. 2.32 The calculate dialogue



3.1416*DBH/2*DBH/2

☑ Variates
□ Factors
□ Texts
□ Scalars
□ Matrices
□ Tables

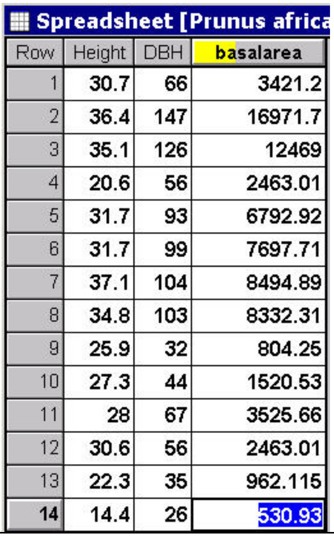
DBH
Height

+	-	*	/	and	eqs
**	*+	(	)	or	nes
<	<=	>	>=	not	is
==	/=	in	ni	eor	isnt

Save Result In: basalarea □ Print in Output

☑ Display In Spreadsheet
○ Display in new sheet
● Add to existing sheet Prunus africana height and dbh Mabira Ugand

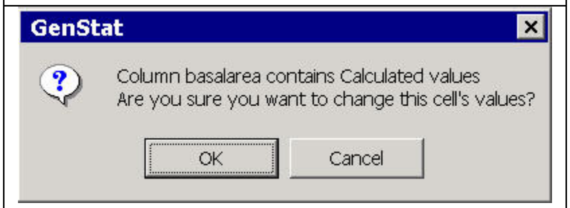
Fig. 2.33 The resulting column


Spreadsheet [Prunus africa

Row	Height	DBH	basalarea
1	30.7	66	3421.2
2	36.4	147	16971.7
3	35.1	126	12469
4	20.6	56	2463.01
5	31.7	93	6792.92
6	31.7	99	7697.71
7	37.1	104	8494.89
8	34.8	103	8332.31
9	25.9	32	804.25
10	27.3	44	1520.53
11	28	67	3525.66
12	30.6	56	2463.01
13	22.3	35	962.115
14	14.4	26	530.93

There is now a new variate, called *basalarea,* added to the spreadsheet, as shown in Fig. 2.33, which holds the 14 values of the basalarea for each tree. The name is part shaded (in yellow on a colour screen) to indicate that the column *basalarea* is a calculated column. To illustrate the difference between an ordinary and a calculated column, try to change a value in the basalarea column. GenStat gives a warning, see Fig. 2.34 below.

Fig. 2.34 Warning if you try to change a value in a calculated column

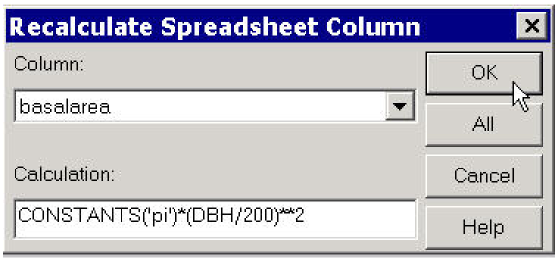
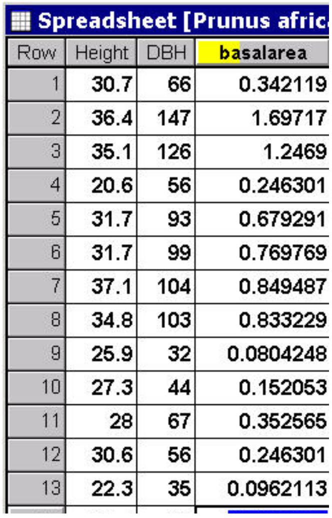


Column basalarea contains Calculated values
Are you sure you want to change this cell's values?

If you are still in the *basalarea* column, right click on the mouse, and choose the option called **Column Attributes**. You will see the Column Attributes dialogue. This gives details of the *basalarea* column, including the calculation you used.

> Thus GenStat's spreadsheet is a little like an ordinary spreadsheet in that it records the calculation, rather than just doing the transformation. If you change a value in the original column, the derived values do not however change automatically. You could then use ***Spread* ⇒ *Calculate* ⇒ *Recalculate***, to update the derived values.

We will do this now, because our calculation contains an error. The diameter of the trees was measured in cm, while usually a basal area is expressed in m^2. So we have to divide each diameter by 200 to get the radius in meters. Meanwhile we can improve the calculation by using the operator ** for the exponent. And instead of abbreviating Pi as 3.1416, we can use the GenStat command for the constant pi: CONSTANTS('pi'). The complete formula is shown in Fig. 2.35.

Fig. 2.35 Correcting the calculation	*Fig. 2.36 The new calculated column*
	

Recalculate Spreadsheet Column

Column: basalarea

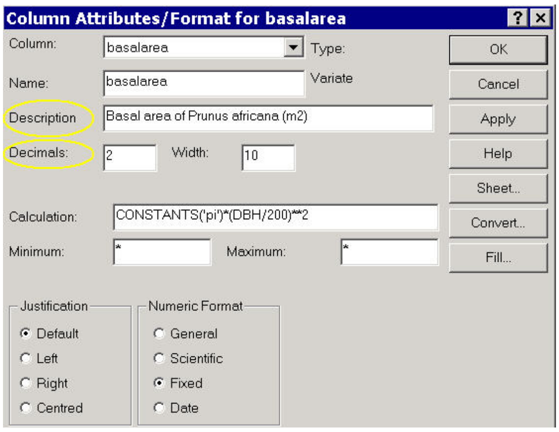
Calculation: CONSTANTS('pi')*(DBH/200)**2

Spreadsheet [Prunus afric

Row	Height	DBH	basalarea
1	30.7	66	0.342119
2	36.4	147	1.69717
3	35.1	126	1.2469
4	20.6	56	0.246301
5	31.7	93	0.679291
6	31.7	99	0.769769
7	37.1	104	0.849487
8	34.8	103	0.833229
9	25.9	32	0.0804248
10	27.3	44	0.152053
11	28	67	0.352565
12	30.6	56	0.246301
13	22.3	35	0.0962113

We have the basal area, but we still could add some descriptive information. Right-click in the *basalarea* column and choose ***Column Attributes*** or choose ***Spread => Column => Attributes/Format*** or press ***[F9]***. Add a description and set the number of decimals to two, as shown in Fig. 2.37.

Fig. 2.37 Description and decimals for the calculated column



Column: basalarea
Type: Variate
Name: basalarea
Description: Basal area of Prunus africana (m2)
Decimals: 2
Width: 10
Calculation: CONSTANTS('pi')*(DBH/200)**2
Minimum: *
Maximum: *

Justification: ● Default ○ Left ○ Right ○ Centred

Numeric Format: ○ General ○ Scientific ● Fixed ○ Date

Now we can calculate the volume of each tree. Choose ***Spread => Calculate => Column*** again or choose ***Window*** and select ***Calculate*** (Fig. 2.38) or click on the window list button in the toolbar (Fig. 2.39). The same window with the first calculation will open.

Fig. 2.38 Getting the calculate dialogue back

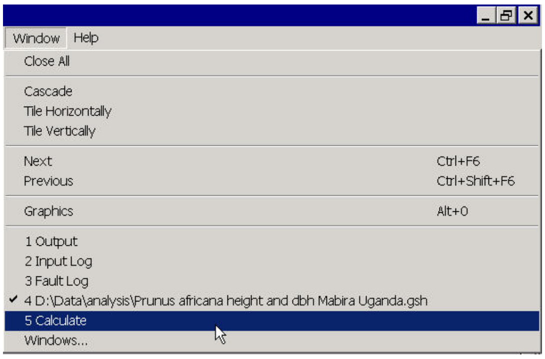
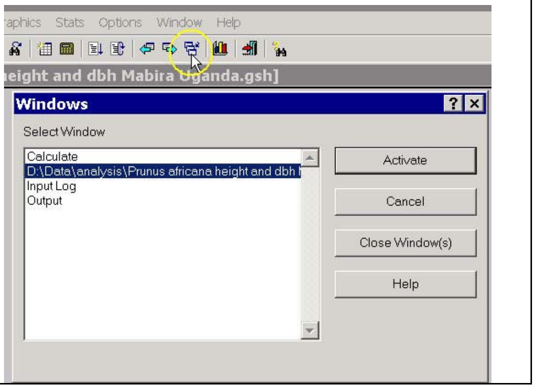


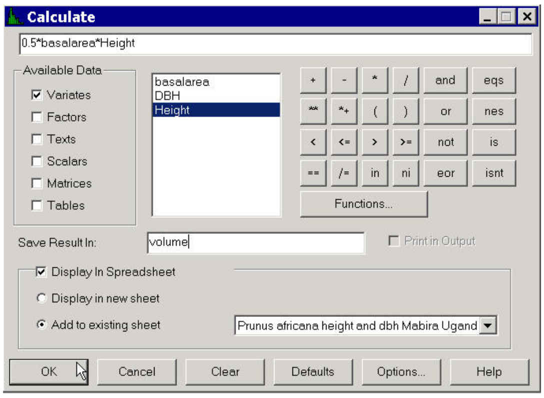
Fig. 2.39 Another way to restore the dialogue


raphics Stats Options Window Help
eight and dbh Mabira Uganda.gsh]
Windows
? ×
Select Window
Calculate
D:\Data\analysis\Prunus africana height and dbh I
Input Log
Output
Activate
Cancel
Close Window(s)
Help

> Many dialogue boxes in GenStat do not close when you click ***[OK]***. They only close if you click on ***[Cancel]***. This is so you can easily repeat an operation, or get more output from the current analysis without having to go back through the menus. It is quite easy to get a large number of windows and dialogue boxes open at once, so it can be quite hard to find the one for which you are looking. Therefore it is a good idea to close a box by clicking ***[Cancel]*** as soon as it is no longer needed.

Calculate the volume and format the column as shown in Fig. 2.40 and Fig. 2.41.

Fig. 2.40 Calculating the volume



0.5*basalarea*Height

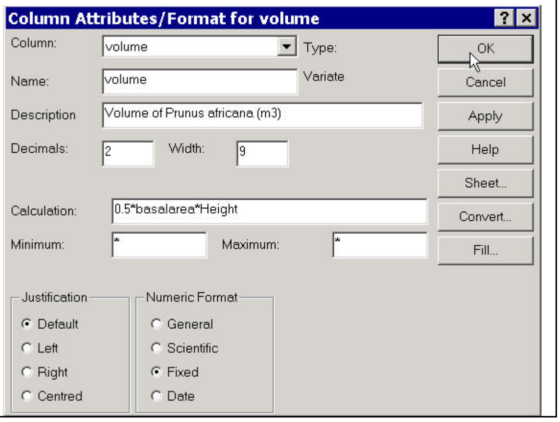
☑ Variates
☐ Factors
☐ Texts
☐ Scalars
☐ Matrices
☐ Tables

basalarea
DBH
Height

Save Result In: volume
☐ Print in Output

☑ Display In Spreadsheet
○ Display in new sheet
● Add to existing sheet: Prunus africana height and dbh Mabira Ugand

Fig. 2.41 Formatting the calculated column



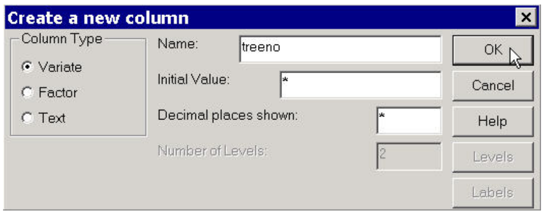
Column: volume
Type: Variate
Name: volume
Description: Volume of Prunus africana (m3)
Decimals: 2
Width: 9
Calculation: 0.5*basalarea*Height
Minimum: *
Maximum: *

Justification
● Default
○ Left
○ Right
○ Centred

Numeric Format
○ General
○ Scientific
● Fixed
○ Date

The data values came from 14 numbered trees. It would be useful to have this information entered too. Click in the first column *(Height)* of the spreadsheet. Choose ***Spread*** ⇒ ***Insert*** ⇒ ***Column before Current Column***. This gives a dialogue box called ***Create a new column*** as shown in Fig. 2.42 below.

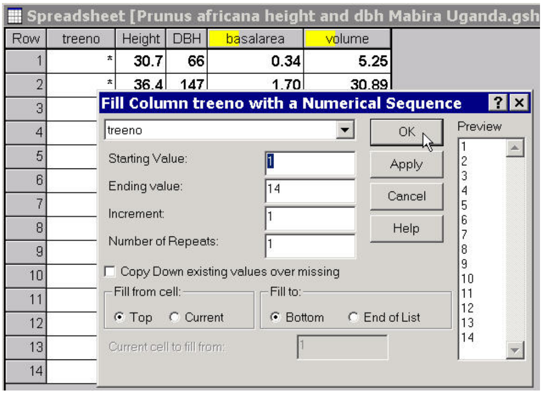
Fig. 2.42 Spread => Insert => Column before Current Column



Column Type
● Variate
○ Factor
○ Text

Name: treeno
Initial Value: *
Decimal places shown: *
Number of Levels: 2

Fig. 2.43 Making a column with a regular sequence


Spreadsheet [Prunus africana height and dbh Mabira Uganda.gsh
Fill Column treeno with a Numerical Sequence
OK
Apply
Cancel
Help
Preview

Row	treeno	Height	DBH	basalarea	volume
1	*	30.7	66	0.34	5.25
2	*	36.4	147	1.70	30.89
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

treeno
Starting Value: 1
Ending value: 14
Increment: 1
Number of Repeats: 1
☐ Copy Down existing values over missing
Fill from cell: ● Top ○ Current
Fill to: ● Bottom ○ End of List
Current cell to fill from: 1

Preview: 1 2 3 4 5 6 7 8 9 10 11 12 13 14

Type *treeno* in the name box and click on ***[OK]***. A new column will appear in the spreadsheet filled with missing values (denoted by *). You could now type in the numbers 1 to 14, but there is a quicker way to fill in regular sequences.

Right click in the Spreadsheet and choose ***Fill*** from the popup menu as shown above or choose ***Spread*** ⇒ ***Calculate*** ⇒ ***Fill***. In the ***Fill*** dialogue that is shown in Fig. 2.43, make sure that *treeno* is in the top box. Clicking ***[OK]*** will fill *treeno* with the numbers 1 to 14. ***Fill*** can also be used to make patterned sequences.

Details of the use of this, or any other dialogue, can be found by clicking the ***[Help]*** button in the dialogue box.

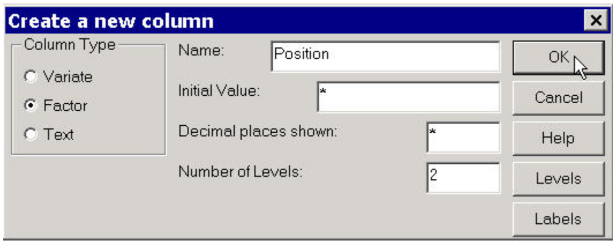
2.3.3 Columns containing factors.

So far, all the information entered into GenStat has been numerical. It is possible to enter textual information as well. One structure that accepts this kind of information is a FACTOR. This is a special column used to indicate groups in the data (there will be more about factors later in this manual).

The first seven trees of this data set were measured in the middle of the forest, the interior, while the last seven trees grew at the forest edge. Hence the factor will have two groups or **levels**. Here one is labelled ***Interior*** and the other ***Edge***.

Click in the first column of the spreadsheet *(rowno)* and choose ***Spread* ⇒ *Insert* ⇒ *Column after Current Column***. Type *Position* into the ***Name*** box, and click to select ***Factor*** under ***Column Type***. The box will change as shown in Fig. 2.44.

Fig. 2.44 Creating a factor column



Create a new column

Column Type: ○ Variate ● Factor ○ Text

Name: Position

Initial Value: *

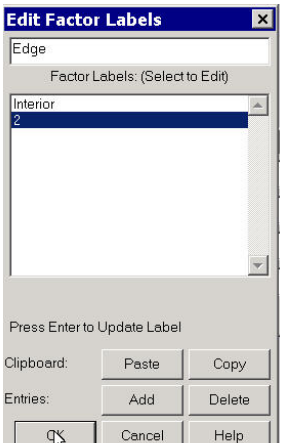
Decimal places shown: *

Number of Levels: 2

Specify that the factor has 2 levels and then click on the ***[Labels]*** button. The dialogue shown below appears. Type '*Interior*' and press the ***[Enter]*** key. The next level (2) will become selected. Type '*Edge*', press ***[Enter]*** and then click ***[OK]*** to make the changes take effect.

Click ***[OK]*** in the ***Create a new column*** dialogue to make the new, column, which contains empty cells (see Fig. 2.46).

Fig. 2.45 Giving labels to factor levels



Edit Factor Labels

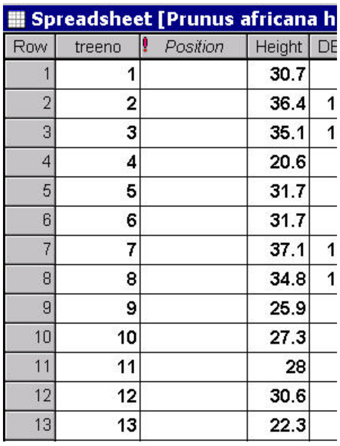
Edge

Factor Labels: (Select to Edit)

Interior
2

Press Enter to Update Label

Fig. 2.46 The resulting spreadsheet



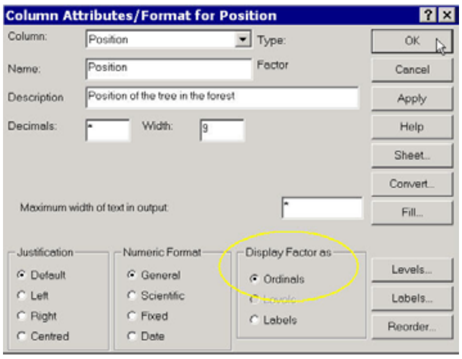
Spreadsheet [Prunus africana h

Row	treeno	Position	Height	DE
1	1		30.7	
2	2		36.4	1
3	3		35.1	1
4	4		20.6	
5	5		31.7	
6	6		31.7	
7	7		37.1	1
8	8		34.8	1
9	9		25.9	
10	10		27.3	
11	11		28	
12	12		30.6	
13	13		22.3	

Now there are two ways of entering the position: entering ordinals or entering labels. Let's do the first 5 trees using the ordinals. Factor values are stored as ordinals; namely as integers between 1 and the number of levels of the factor. In our example there are two factor levels, so the ordinals will be 1,1,1,1,1,1,1,2,2,2,2,2,2,2.

Right-click in the empty Position column and choose ***Column Attributes***. Indicate that the factor has to be displayed as ordinals (Fig. 2.47). Now enter 1 for the first 5 trees, see Fig. 2.48.

Fig. 2.47 Displaying a factor as ordinals

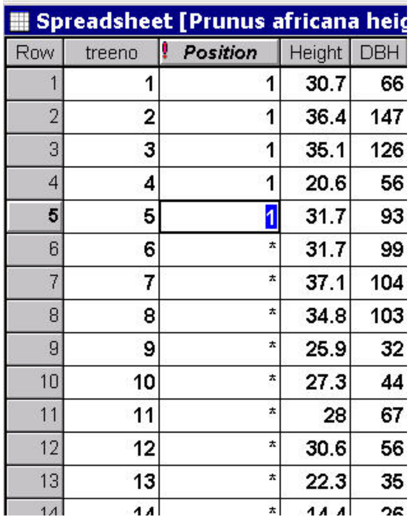


Column Attributes/Format for Position

Column: Position
Type: Factor
Name: Position
Description: Position of the tree in the forest
Decimals: *
Width: 9
Maximum width of text in output: *

Justification: ● Default ○ Left ○ Right ○ Centred
Numeric Format: ● General ○ Scientific ○ Fixed ○ Date
Display Factor as: ● Ordinals ○ Levels ○ Labels

Fig. 2.48 Entering the first level of a factor



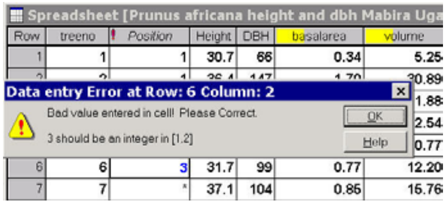
Spreadsheet [Prunus africana heig

Row	treeno	Position	Height	DBH
1	1	1	30.7	66
2	2	1	36.4	147
3	3	1	35.1	126
4	4	1	20.6	56
5	5	1	31.7	93
6	6	*	31.7	99
7	7	*	37.1	104
8	8	*	34.8	103
9	9	*	25.9	32
10	10	*	27.3	44
11	11	*	28	67
12	12	*	30.6	56
13	13	*	22.3	35

Try to enter 3 as Position for tree number 6. GenStat gives a warning that only 1 or 2 is possible (see Fig. 2.49). Click ***[OK]*** and press ***[Escape]***.

Right-click again in the Position column and choose ***Column Attributes*** to give the dialogue shown in Fig. 2.47. This time however, indicate that the factor has to be displayed as labels. We had entered the factor labels already, so after clicking ***[OK]***, the position of the first five trees is shown as *Interior*. Now you can continue entering the values. You can enter 'Interior', 'interior' or even just the first letter 'i' and GenStat will show the correct factor label 'Interior'. Enter 'e' for trees 8 until 14 as shown in Fig. 2.50.

Fig. 2.49 Attempt to enter an illegal value in a factor column



Spreadsheet [Prunus africana height and dbh Mabira Uga

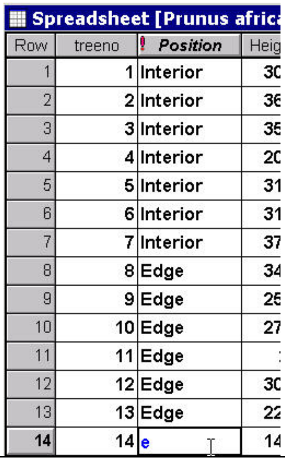
Row	treeno	Position	Height	DBH	basalarea	volume
1	1	1	30.7	66	0.34	5.25
6	6	3	31.7	99	0.77	12.20
7	7	*	37.1	104	0.85	15.76

Data entry Error at Row: 6 Column: 2

Bad value entered in cell! Please Correct.

3 should be an integer in [1.2]

Fig. 2.50 Entry of data into a factor column using the labels

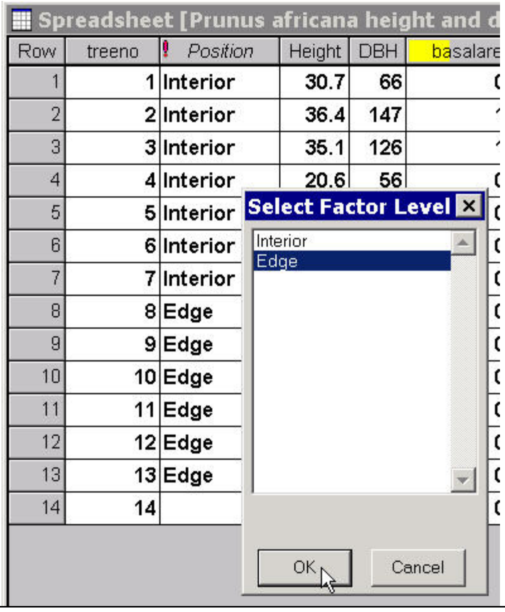


Spreadsheet [Prunus afric

Row	treeno	Position	Heig
1	1	Interior	30
2	2	Interior	36
3	3	Interior	35
4	4	Interior	20
5	5	Interior	31
6	6	Interior	31
7	7	Interior	37
8	8	Edge	34
9	9	Edge	25
10	10	Edge	27
11	11	Edge	
12	12	Edge	30
13	13	Edge	22
14	14	e	14

As long as you type the right first letter of the Factor label, GenStat will display the correct label. If you type the wrong letter, GenStat will give you a message and ask you to retype your entry. Double clicking gives a pop-up menu that lists the allowable levels; see Fig. 2.51.

Fig. 2.51 Popup menu to indicate the allowed factor labels



Spreadsheet [Prunus africana height and d

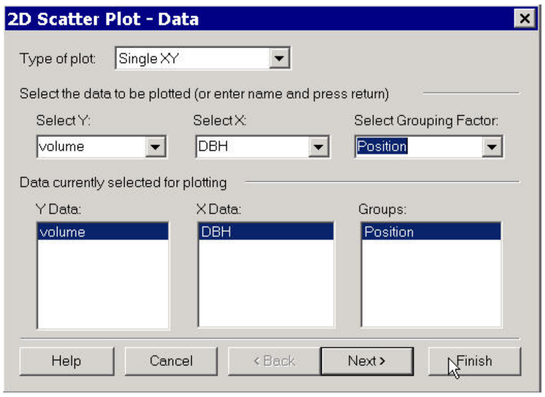
Row	treeno	Position	Height	DBH	basalare
1	1	Interior	30.7	66	
2	2	Interior	36.4	147	
3	3	Interior	35.1	126	
4	4	Interior	20.6	56	
5	5	Interior			
6	6	Interior			
7	7	Interior			
8	8	Edge			
9	9	Edge			
10	10	Edge			
11	11	Edge			
12	12	Edge			
13	13	Edge			
14	14				

Select Factor Level

Interior
Edge

The *Position* column can be used to label a graph. Choose ***Graphics ⇒ Point Plot => Single XY type***. Fill in the boxes as below, and click ***[Finish]***. If you first click ***[Next] in*** Fig. 2.52, you can add titles to the graph and the axes.

Fig. 2.52 Graphics => Point Plot => Single XY



2D Scatter Plot - Data

Type of plot: Single XY

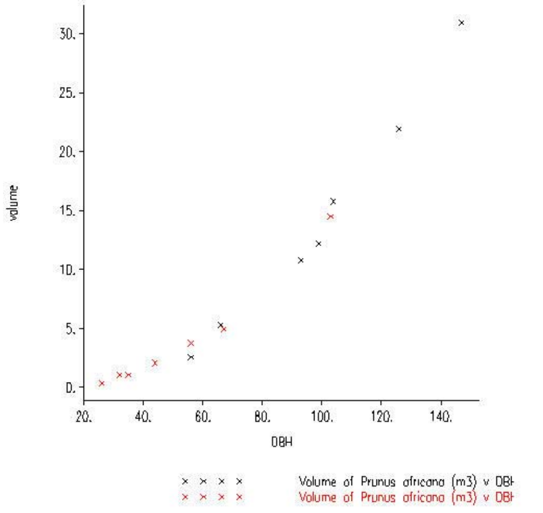
Select the data to be plotted (or enter name and press return)

Select Y: volume
Select X: DBH
Select Grouping Factor: Position

Data currently selected for plotting

Y Data: volume
X Data: DBH
Groups: Position

Fig 2.53 Colours in graph for different levels



In the screen plot, the points from the two groups will be coloured differently, but both will be plotted as X.

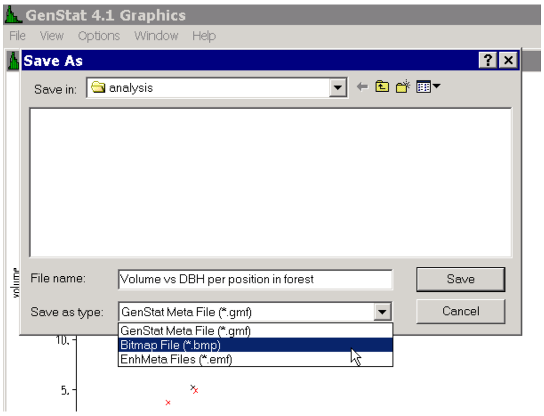
In the GenStat Discovery Edition (based on GenStat for Windows 5), only an older version of the graphics editor is available (GenStat 4.1 Graphics). You can add a general title and a title to the X and Y axis, add an arrow to the axes and change the tick marks. But that’s basically it. In GenStat for Windows from version 5 SP2, also a new graphics editor is available, with many more possibilities (changing an existing graph, different symbols and colours, zooming, rotating, many more file formats, …). In chapter 6.3 we will show some ways to work around the limitations of the GenStat 4.1 Graphics editor.

In the Discovery Edition, graphs can be saved in 3 different formats by choosing ***File => Save as***:

- *.**gmf** – GenStat Meta File. This is the default GenStat graphics format. You can reopen a GenStat Meta File in GenStat and send it to other GenStat users. You will not be able to insert a gmf file as a picture in MS Word.
- *.**bmp** – Bitmap File. In this file format, graphics are stored as pixels. It can only be used on the Windows platform. File compression is not supported so bmp files are usually large.
- *.**emf** – Enhanced Meta File. This is another graphics file format for the Windows platform, the successor of wmf (Windows Meta File) format. In the Meta file format, graphics can be stored both as bitmap (pixels) or as vector format (commands like “draw line”). Emf is only supported on Windows 95 and higher. Not all software supports emf, but MS Word 97 or later can import it.

If you want to create a temporary graph that you will only use in GenStat, choose the *.gmf format. If you want to include a picture in a Word document, choose the *.emf format. For other purposes, use the bmp format. In the graphics editor you can change the pixel size (***Options => Change Bitmap Size***). If you want to make really impressive graphics, it’s however better to use GenStat version 7 or another software.

Fig. 2.54 Saving a graph as bitmap file



Save in: analysis

File name: Volume vs DBH per position in forest

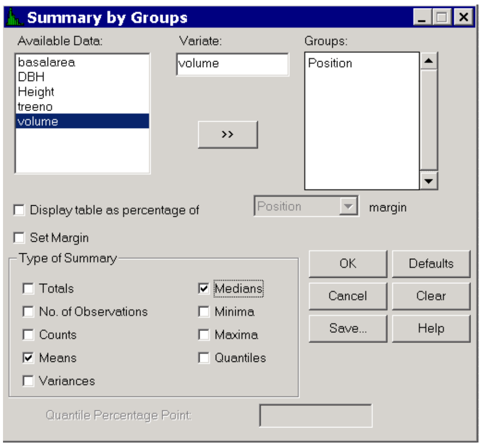
Save as type: GenStat Meta File (*.gmf)

GenStat Meta File (*.gmf)
Bitmap File (*.bmp)
EnhMeta Files (*.emf)

You leave the GenStat Graphics Window by choosing ***File ⇒ Exit*** from the menu bar.

Earlier, you used ***Stats ⇒ Summary Statistics ⇒ Summarise Contents of Variates*** to give some summaries of the data. Now, with the data in two groups, it is useful to give the summaries for each group individually. The dialogue used in Fig. 2.25 can be used for this, but a more general alternative is ***Stats ⇒ Summary Statistics ⇒ Summaries of Groups (Tabulation)*** to give the dialogue shown in Fig. 2.55.

Fig. 2.55 The tabulation dialogue



Summary by Groups

Available Data: basalarea, DBH, Height, treeno, volume

Variate: volume

Groups: Position

☐ Display table as percentage of Position margin

☐ Set Margin

Type of Summary

☐ Totals ☑ Medians

☐ No. of Observations ☐ Minima

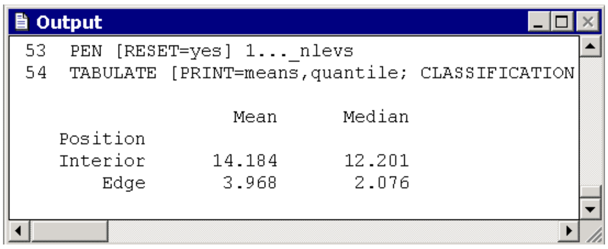
☐ Counts ☐ Maxima

☑ Means ☐ Quantiles

☐ Variances

Quantile Percentage Point:

Fig. 2.56 Summary statistics for each factor level



Output

```
 53  PEN [RESET=yes] 1..._nlevs
 54  TABULATE [PRINT=means,quantile; CLASSIFICATION

                      Mean       Median
    Position
    Interior        14.184       12.201
        Edge         3.968        2.076
```

Complete the dialogue as shown and press ***[OK]***. The results as shown in Fig. 2.56 appear in the Output Window.

Save the spreadsheet before continuing

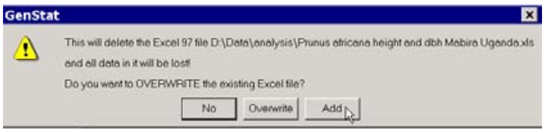
2.3.4 Saving data from GenStat to Excel.

In chapter 2.2.1.3 on page 10 we saw already how to save a spreadsheet. By default, a Window appears asking to save the data as a ***GenStat spreadsheet*** (*.gsh). This is particularly useful if backward compatibility with older GenStat versions is wanted. But a wide range of other file formats is also available.

In chapter 2.2.2 it was shown how data could be imported from an Excel worksheet. We had imported the file "*Prunus africana height and dbh Mabira Uganda.xls*" from Excel. We will reopen this file and calculate the basal area again. Since we did all calculations in the GenStat spreadsheet "*Prunus africana height and dbh Mabira Uganda.ghs*", the Excel spreadsheet only contains 2 colums Height and DBH. So, first choose ***Run => Restart Session***, indicating ***[Yes]*** to clear all data from the GenStat memory, open the Excel file and calculate the basal area. Refer to chapter 2.3.2 if necessary.

This time we want to save the spreadsheet as an Excel file. Choose ***File => Save***.

Fig. 2.57 Add to the Excel file

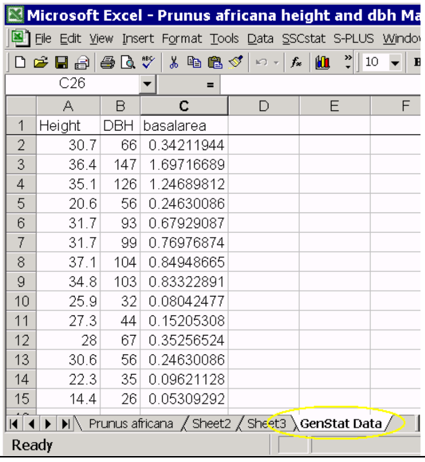


GenStat

This will delete the Excel 97 file D:\Data\analysis\Prunus africana height and dbh Mabira Uganda.xls and all data in it will be lost!

Do you want to OVERWRITE the existing Excel file?

Fig. 2.58 A new Excel sheet is added



Microsoft Excel - Prunus africana height and dbh Ma

File Edit View Insert Format Tools Data SSCstat S-PLUS Window
C26
=
Prunus africana
Sheet2
Sheet3
GenStat Data
Ready

	A	B	C	D	E	F
1	Height	DBH	basalarea			
2	30.7	66	0.34211944			
3	36.4	147	1.69716689			
4	35.1	126	1.24689812			
5	20.6	56	0.24630086			
6	31.7	93	0.67929087			
7	31.7	99	0.76976874			
8	37.1	104	0.84948665			
9	34.8	103	0.83322891			
10	25.9	32	0.08042477			
11	27.3	44	0.15205308			
12	28	67	0.35256524			
13	30.6	56	0.24630086			
14	22.3	35	0.09621128			
15	14.4	26	0.05309292			

The result is a warning message, shown in Fig. 2.57. When you click ***[Overwrite]***, all existing worksheets in the workbook "*Prunus africana height and dbh Mabira*

Uganda.xls" will be deleted and the data will be saved in a worksheet called ***GenStat Data***. By clicking ***[Add]***, the existing worksheets will be kept and a new worksheet ***GenStat Data*** will be added to the Excel file, Fig. 2.58. If you repeat this, new worksheets will be added: GenStat Data, GenStat Datb, GenStat Datc, …

2.3.5 Importing factors from Excel.

If you import data from Excel that contain factors, they are treated slightly differently. In our example, *Interior* was the first factor level or ordinal and *Edge* was the second. If you import a column containing the factors "*Interior*" and "*Edge*" from Excel, *Edge* would have an ordinal of 1 and *Interior* an ordinal of 2. The reason is that Excel reads factors from Excel in alphabetical order.

2.3.6 Deleting data.

Before proceeding we delete the column, called *treeno* to show the difference between deleting a whole column and deleting its contents.

First we select the column. Click in the name field; click in the column and press ***[Alt]+[Ctrl]+C***; or choose ***Spread ⇒ Select ⇒ Current Column***. Clicking again will deselect the column. Once selected, you might think that the ***[Delete]*** key should delete the column. If you press the ***[Delete]*** key, only the data disappear, the column remains! Use ***Edit ⇒ Undo Del Cells*** or press ***[Ctrl] + Z*** to get the data back (Fig. 2.59).

To delete the whole column, with the cursor in the column choose ***Spread ⇒ Delete ⇒ Current Column***. You still can recover the column choosing ***Edit ⇒ Undo Del Col*** or by pressing ***[Ctrl] + Z*** (Fig. 2.60). You can also select one, or more, rows and delete them in the same way.

Fig. 2.59 Undoing the deletion of cells

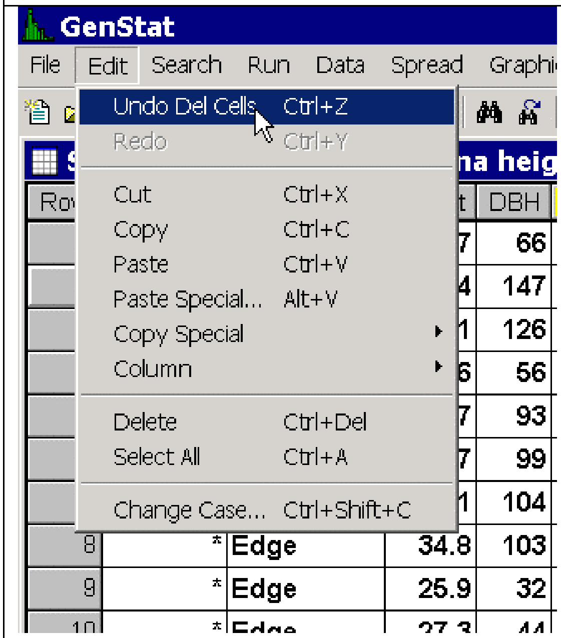
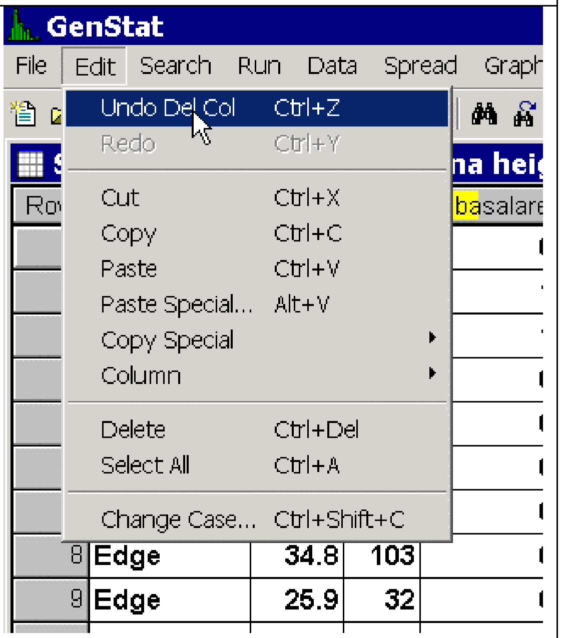


Fig. 2.60 Undoing the deletion of columns

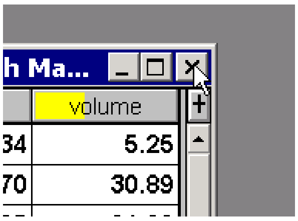


2.4 Understanding how GenStat works.

2.4.1 Available variables.

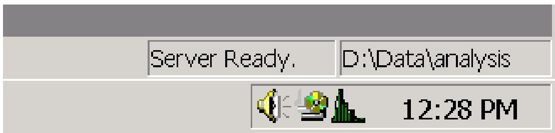
Close the spreadsheet with the Prunus africana data. You can do this by choosing ***File => Close***, by pressing ***[Ctrl]+[F4]*** or by clicking on the button with a diagonal cross in the top right hand corner of the spreadsheet. Once closed, do you think the data are still in GenStat?

Fig. 2.61 Closing a GenStat spreadsheet



The answer is yes. Because the GenStat you see is a Windows interface that sends commands to a program running in the background: the Genstat Server. When these commands are processed, the message in the GenStat status bar shows what is happening and the GenStat icon in the Windows Taskbar changes from green (Fig. 2.62) to red (Fig. 2.63). However, when working on small datasets it goes so fast that you will not be able to see it.

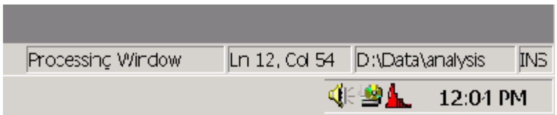
Fig. 2.62 The taskbar with the GenStat server ready



Server Ready. D:\Data\analysis

12:28 PM

Fig. 2.63 The GenStat server icon changes to red when the server is busy

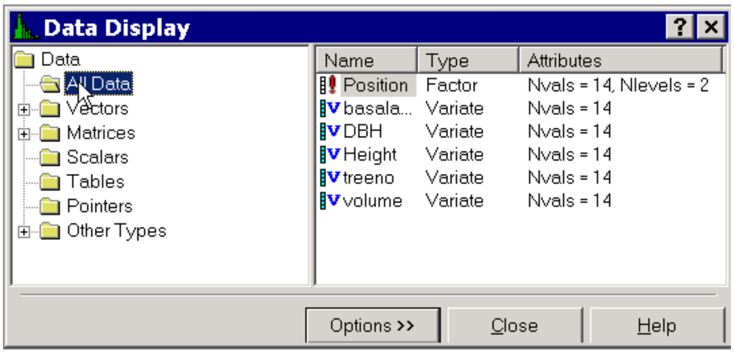


Processing Window Ln 12, Col 54 D:\Data\analysis INS

12:01 PM

So, even if you don't see anything, there can be still all kinds of data somewhere in the GenStat server. You can check which variables are currently available to the GenStat server using ***Data ⇒ Display*** or pressing the ***[F5]*** key and clicking for instance on "All Data".

Fig. 2.64 List of all the variables in the GenStat server



Name	Type	Attributes
Position	Factor	Nvals = 14, Nlevels = 2
basala...	Variate	Nvals = 14
DBH	Variate	Nvals = 14
Height	Variate	Nvals = 14
treeno	Variate	Nvals = 14
volume	Variate	Nvals = 14

> This lists the names of the structures and their **types** as shown in Fig. 2.64. All structures used so far are **variates** (Height, DBH, basalarea, volume, treeno) and **factors** (Position), but later on you will use other types of columns too. This is also a useful dialogue box from which you can delete columns when they are no longer needed.

Click ***[Close]*** to close the ***Display*** dialogue box. See chapter 2.4.3 for information on how to clean the GenStat Server of data.

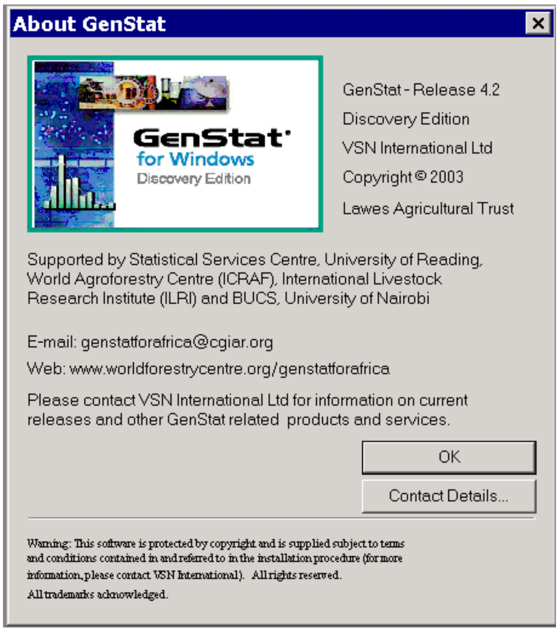
2.4.2 A first introduction to the GenStat command language.

So, GenStat is basically a standard Windows application running on top of the GenStat server. GenStat existed long before Windows was created and in the old days you had to know the "language". You simply typed commands, which you submitted to GenStat.

The menus in the GenStat Discovery Edition are based on an underlying command language, ‘GenStat release 4.2’, see Fig. 2.65. Release 4.2 means it is based on the 4th major revision of the GenStat Server that has undergone 2 minor revisions. The Discovery Edition itself is based on a slightly modified GenStat for Windows fifth edition.

You can still use GenStat by typing commands in the Input Window as we show now. At the same time, we show how GenStat is used as a calculator.

Fig. 2.65 Details about GenStat



GenStat - Release 4.2
Discovery Edition
VSN International Ltd
Copyright © 2003
Lawes Agricultural Trust

Supported by Statistical Services Centre, University of Reading, World Agroforestry Centre (ICRAF), International Livestock Research Institute (ILRI) and BUCS, University of Nairobi

E-mail: genstatforafrica@cgiar.org

Web: www.worldforestrycentre.org/genstatforafrica

Please contact VSN International Ltd for information on current releases and other GenStat related products and services.

Warning: This software is protected by copyright and is supplied subject to terms and conditions contained in and referred to in the installation procedure (for more information, please contact VSN International). All rights reserved.

All trademarks acknowledged.

Restart GenStat. Use ***File ⇒ New ⇒ Text Window*** as shown in Fig. 2.66. This gives you an ***Input Window***. In this window, type ***Print 3+4*** as shown in Fig. 2.67.

Fig. 2.66 Opening a text window

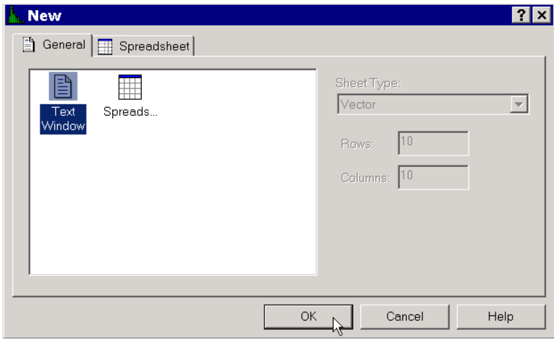
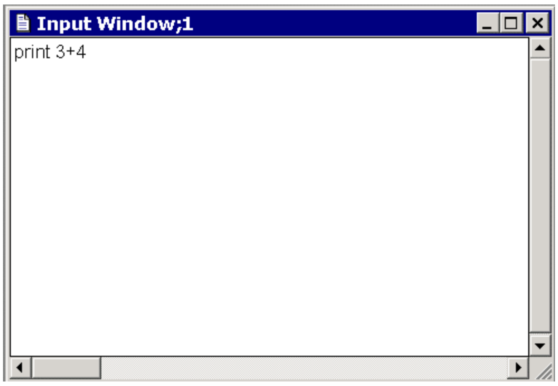


Fig. 2.67 Typing a GenStat command



```
print 3+4
```

Now select the ***Run*** menu (Fig. 2.68). You can choose either ***Submit Line*** (if the cursor is still on the line you typed) or ***Submit Window***. Choose one of these.

Fig. 2.68 Submitting commands to GenStat

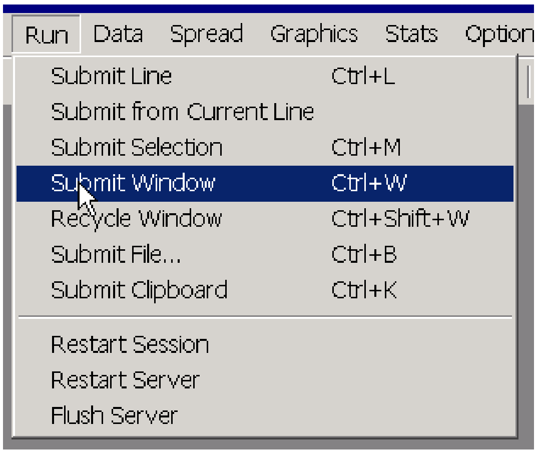
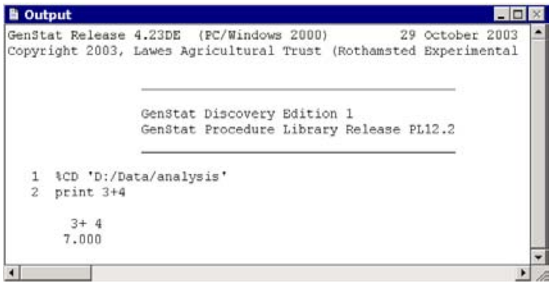


Fig. 2.69 The results are in the output window



```
GenStat Release 4.23DE  (PC/Windows 2000)         29 October 2003
Copyright 2003, Lawes Agricultural Trust (Rothamsted Experimental

                  ________________________________________

                  GenStat Discovery Edition 1
                  GenStat Procedure Library Release PL12.2
                  ________________________________________

   1  %CD 'D:/Data/analysis'
   2  print 3+4

      3+ 4
     7.000
```

You have now submitted your "program" of commands to the GenStat server. The results are put in the ***Output Window***.

You can go to the output window in various ways, e.g. by using the Windows menu. There you see that GenStat normally "echoes" the command and shows you that 3+4=7.

The Windows version of GenStat gives you a variety of ways of submitting calculations to the GenStat server. An alternative to the above is to use the Data menu: ***Data*** ⇒ ***Calculations*** as shown in Fig. 2.70.

Fig. 2.70 Data => Calculations

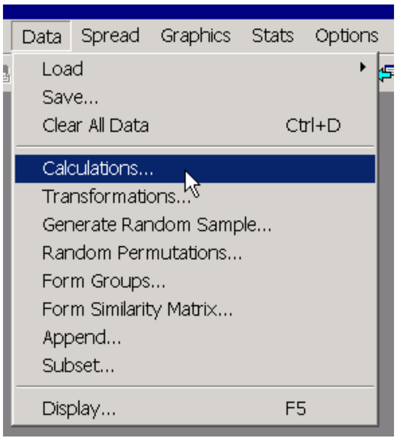
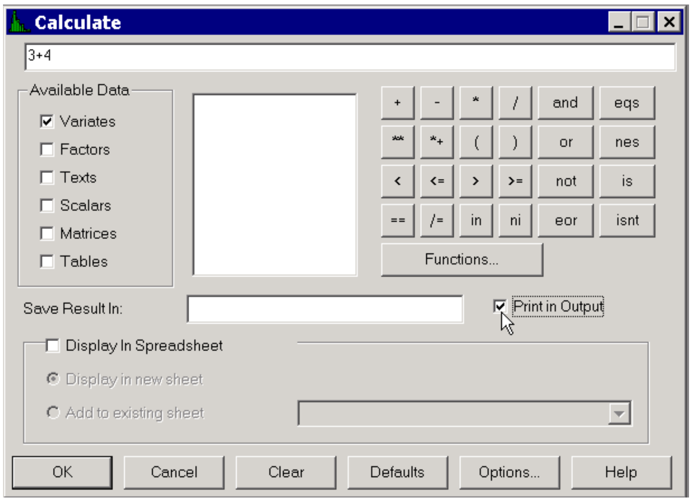


Fig. 2.71 Using the calculate dialogue



3+4

☑ Variates
□ Factors
□ Texts
□ Scalars
□ Matrices
□ Tables

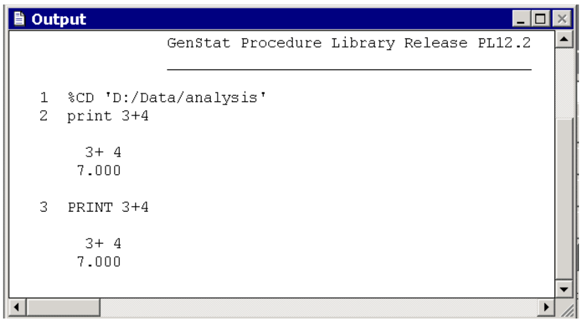
Save Result In:

☑ Print in Output

□ Display In Spreadsheet

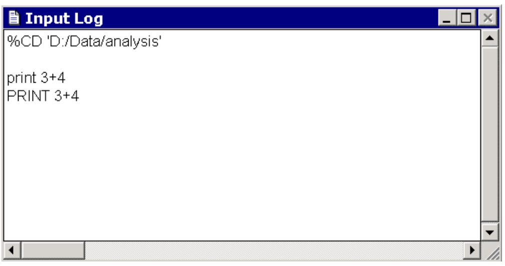
Then type ***3 + 4*** as the function, click on ***Print in Output*** and then on ***[OK]***. If you look in the Output window, you see that 3 + 4 still equals 7 (Fig. 2.72).

Fig. 2.72 And yes: 3 + 4 still is 7



```
                  GenStat Procedure Library Release PL12.2
                  ________________________________________

   1  %CD 'D:/Data/analysis'
   2  print 3+4

        3+ 4
       7.000

   3  PRINT 3+4

        3+ 4
       7.000
```

Fig. 2.73 The input log



```
%CD 'D:/Data/analysis'

print 3+4
PRINT 3+4
```

The ***Input Log Window*** is also useful. It keeps a record of all the commands you have submitted, see Fig. 2.73. Access it for instance by choosing ***Window ⇒ Input Log***. You see that the use of the **Calculation** menu has resulted in GenStat preparing the commands `PRINT 3+4` for you and has submitted them to the GenStat server.

So, that is how GenStat works. You prepare commands, which are submitted to the GenStat server. The Windows version has simply given you a variety of ways of helping you prepare the commands for GenStat. GenStat obeys the commands and puts the results in the ***Output Window***. It keeps a record in the ***Input Window***.

If the commands produce graphs, then GenStat puts the graphs in a ***Graphics Window***. If you make a mistake in the command, it prints an error message in the ***Fault Window*** (and in the ***Output Window***).

The example above (3 + 4 = 7) indicates that GenStat may be used as a simple calculator. This is worth a little practice. It is useful to have a scientific calculator. Also it is sometimes useful to transform data. For example, if you want to calculate the difference between 4.35 and 2.37 expressed as a percentage of 4.35, open the calculator with ***Data ⇒ Calculations***, check that ***Print in Output***, is still ticked and type the following calculation in the top box:

```
100 * (4.35 -2.37) / 4.35
```

Click ***[OK]***. This will give the following in the output window:

```
            (100*(4.35- 2.37))/ 4.35
                        45.52
```

i.e. the difference is 45.52% of 4.35.

It is important that the brackets () are included where appropriate to make sure that the calculation has only one meaning.

Try more calculations to see how this works, using both an ***Input window*** and the ***Data ⇒ Calculations*** dialogue box.

The symbols +, -, *, / are used for the operations of addition, subtraction, multiplication and division respectively and ** is used for powers. There are also various mathematical functions available. One is for calculating the square root of a number. The function is ***SQRT()***, where the number whose square root is required is given in the parenthesis, for example SQRT(12.37). The following table gives an overview of how to perform some calculations by using the Input Window. More information can be found in the GenStat Help file under '*List of functions for expressions*'.

Some basic calculations using the Input Window			
Symbol	**Operation**	**Example**	**Result**
+	addition	PRINT 3+4	7.000
-	subtraction	PRINT 3-4	- 1.000
*	product	PRINT 3*4	12.00
/	division	PRINT 3/4	0.7500
**	exponentiation	PRINT 3**4	81.00
Function	**Operation**	**Example**	**Result**
SQRT(x)	Square root	PRINT SQRT(4)	2.00
EXP(x)	Exponential function	PRINT EXP(1)	2.718
LOG(x)	natural logarithm of x, for x > 0	PRINT LOG(2.718)	0.9999
LOG10(x)	logarithm to base 10 of x, for x > 0.	PRINT LOG10(10)	1.000
ROUND(x)	rounds the values of x to the nearest integer.	PRINT ROUND(1.234)	1.000
Other examples			**Result**
PRINT (1/2)			0.5000
PRINT (100*(4.35 -2.37))/4.35			45.52
PRINT CONSTANTS('pi')			3.142
PRINT CONSTANTS('e')			2.718

By default, GenStat will show three decimals in the Output Window when using the PRINT command, or **PRINT directive** in the GenStat terminology. To increase this you have to add a parameter to this directive.

PRINT CONSTANTS(‘pi’); DECIMALS = 10

will give you 3.141592654 in the Output Window.

Most of the time however, you will perform calculations in a spreadsheet as was seen in chapter 2.3.2 above. Once you become experienced in using GenStat, you could alternatively do calculations only in the GenStat server, using the ***Data ⇒ Calculations*** menu, rather than the ***Spread ⇒ Calculate ⇒ Column*** route that you used earlier. The result is the same to the GenStat Server, but you would not automatically see the calculated column in a spreadsheet.

2.4.3 Server sessions.

After trying several of the above calculations, the Input and Output Windows look a mess. All the data can be cleared out of the GenStat server with ***Data ⇒ Clear All Data*** or ***Run ⇒ Restart Session***. Less drastically, you can clear the output window by clicking the ‘Clear Output’ button (☒) in the toolbar.

Input as well as Output Windows can be saved (make the window active by clicking in it and click ***File => Save As***). You can save the Input Window as a text file or as a ‘GenStat file” (*.gen). In this way, you can load the commands in the Input Window again for another, similar, analysis. The Output Window can be saved as a text file or as an ‘Output file’ (*.out). This is useful to save the results of an analysis for comparison with other results or for reporting. Saving a selection of input and output files of critical analyses also contributes in establishing an audit trail.

3 Simple statistical ideas

The main purpose of chapter 2 was to introduce GenStat. Chapter 2 can be considered as a tutorial. In this and the following chapters we still show how GenStat Discovery Edition operates but we also review some basic statistics and show how GenStat can be used to teach statistics. Most of the examples are taken from Mead, Curnow and Hasted[1]. For more information on the statistical aspect of the examples, see the relevant section in that book or refer to any other introductory text.

3.1 First some more data manipulation: appending spreadsheets

In the analysis so far, we have just considered descriptive statistics. Thus we have summarised the data numerically and drawn graphs. In the next chapters we introduce ideas of simple statistical inference. However, first we introduce some more data manipulation.

We take an example from *Mead, Curnow and Hasted*, pages 36 and 42. This compares 6 observations from a new variety of wheat that have following yields (tons/ha):

new: **2.5 2.1 2.4 2.0 2.6 2.3**

with 10 observations from the standard variety:

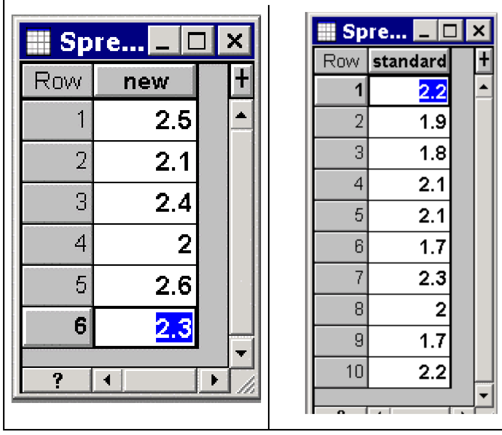
standard: **2.2 1.9 1.8 2.1 2.1 1.7 2.3 2.0 1.7 2.2**

Because these columns are of different lengths, they are entered into two separate spreadsheets. For the first set, use ***Spread* ⇒ *New* ⇒ *Blank*** as shown earlier in chapter 2.2.1 (page 7). Set it to have 1 column of 6 rows, enter the data and give the column the name *new*.

Save the spreadsheet, giving it a meaningful name as was seen in chapter 2.2.1.3 on page 10, for instance "*Wheat variety new.gsh*". Then use ***Spread* ⇒ *New* ⇒ *Blank*** again. Change the number of rows to 10 and enter the second set of data into this other spreadsheet, naming the column as *standard*. Save the spreadsheet, giving it another name, for instance "*Wheat variety standard.gsh*" (see Fig. 3.2).

[1] Roger Mead, Robert N. Curnow, Anne M. Hasted, 2003. Statistical Methods in Agriculture and Experimental Biology. Third Edition. Chapman & Hall/CRC. 472 pages ISBN 1-58488-187-9

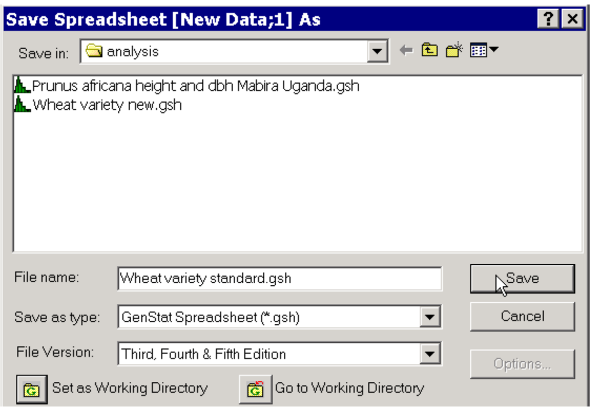
Fig. 3.1 Spreadsheets for wheat yields



Row	new
1	2.5
2	2.1
3	2.4
4	2
5	2.6
6	2.3

Row	standard
1	2.2
2	1.9
3	1.8
4	2.1
5	2.1
6	1.7
7	2.3
8	2
9	1.7
10	2.2

Fig. 3.2 Naming the spreadsheets



Save Spreadsheet [New Data;1] As

Save in: analysis

Prunus africana height and dbh Mabira Uganda.gsh
Wheat variety new.gsh

File name: Wheat variety standard.gsh

Save as type: GenStat Spreadsheet (*.gsh)

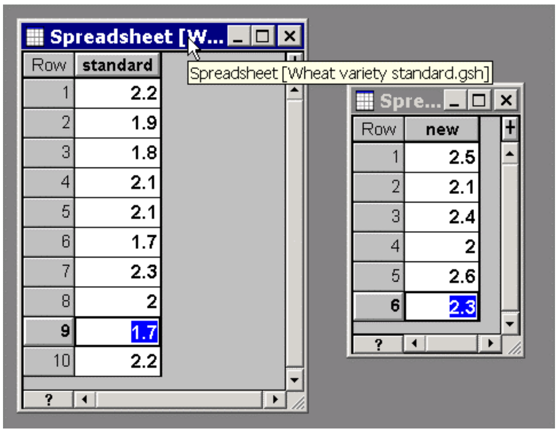
File Version: Third, Fourth & Fifth Edition

Data often need reorganising before analysis and here this step is illustrated by joining the data together from the two sets.

What we wish to do is to append the data from the two columns and add a further column that specifies from which set each observation has come.

If the spreadsheets are no longer in GenStat then they will have to be opened. They were saved earlier with the names '*Wheat variety new.gsh*' and *'Wheat variety standard.gsh' (*Fig. 3.1).

Fig. 3.3 The spreadsheet 'Wheat variety standard.gsh' is the active window.

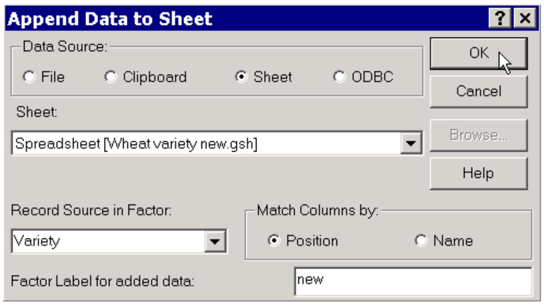


Spreadsheet [Wheat variety standard.gsh]

Row	standard
1	2.2
2	1.9
3	1.8
4	2.1
5	2.1
6	1.7
7	2.3
8	2
9	1.7
10	2.2

Row	new
1	2.5
2	2.1
3	2.4
4	2
5	2.6
6	2.3

Fig. 3.4 The append dialogue box.



Append Data to Sheet

Data Source: ○ File ○ Clipboard ● Sheet ○ ODBC

Sheet: Spreadsheet [Wheat variety new.gsh]

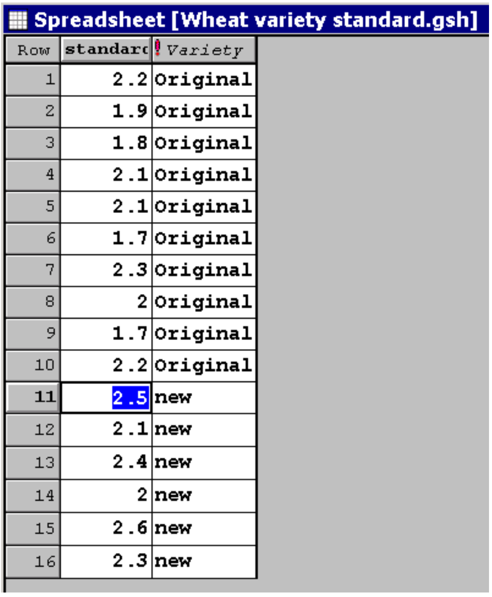
Record Source in Factor: Variety

Match Columns by: ● Position ○ Name

Factor Label for added data: new

Click in the spreadsheet *Wheat variety standard.gsh*, so it is the active window (Fig. 3.3). Use ***Spread*** ⇒ ***Manipulate*** ⇒ ***Append*** and complete the dialogue as shown in Fig. 3.4, i.e. append *Wheat variety new.gsh* to the data in *Wheat variety standard.gsh.* We also specify that a factor column with the name *variety* will be used to distinguish between the two sets of data, and that the second level of the factor will get the name '*new*'. Press ***[OK]***.

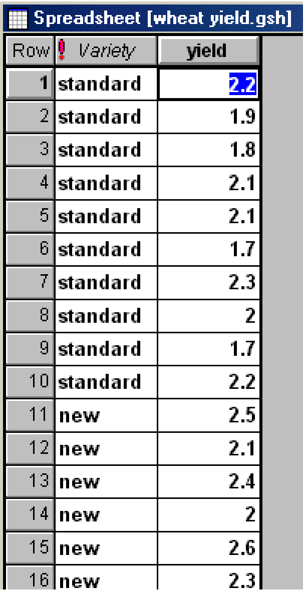
Fig. 3.5 The resulting spreadsheet after the append operation.



Spreadsheet [Wheat variety standard.gsh]

Row	standarc	! Variety
1	2.2	Original
2	1.9	Original
3	1.8	Original
4	2.1	Original
5	2.1	Original
6	1.7	Original
7	2.3	Original
8	2	Original
9	1.7	Original
10	2.2	Original
11	2.5	new
12	2.1	new
13	2.4	new
14	2	new
15	2.6	new
16	2.3	new

Fig. 3.6 The final spreadsheet after some renaming.



Spreadsheet [wheat yield.gsh]

Row	! Variety	yield
1	standard	2.2
2	standard	1.9
3	standard	1.8
4	standard	2.1
5	standard	2.1
6	standard	1.7
7	standard	2.3
8	standard	2
9	standard	1.7
10	standard	2.2
11	new	2.5
12	new	2.1
13	new	2.4
14	new	2
15	new	2.6
16	new	2.3

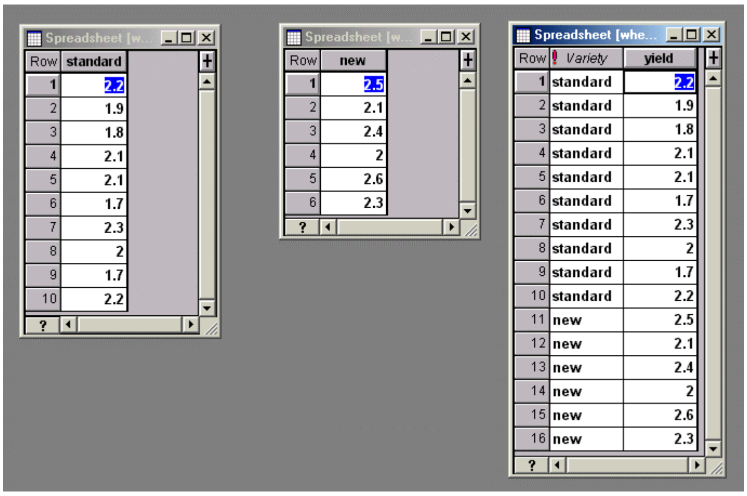
The resulting spreadsheet is as shown in Fig. 3.5 above. This latter form of the data is more common and is used in most of the remainder of this guide.

Now it is just a matter of cleaning up to get the spreadsheet as in Fig. 3.6

- Change the label of the first factor level from *Original* to *standard*. (see)
- Rename the column with the variables from *standard* to *yield*. (see chapter 2.2.1.1)
- Save the spreadsheet as '*wheat yield.gsh*'. (see chapter 2.2.1.3)

So, you have the data in the GenStat server shown in Fig. 3.7 and Fig. 3.8.

Fig. 3.7 The visible spreadsheets



Spreadsheet [w...

Row	standard
1	2.2
2	1.9
3	1.8
4	2.1
5	2.1
6	1.7
7	2.3
8	2
9	1.7
10	2.2

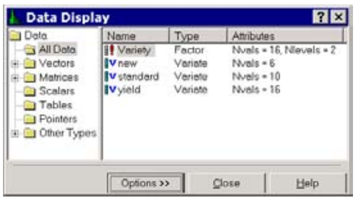
Spreadsheet [w...

Row	new
1	2.5
2	2.1
3	2.4
4	2
5	2.6
6	2.3

Spreadsheet [whe...

Row	! Variety	yield
1	standard	2.2
2	standard	1.9
3	standard	1.8
4	standard	2.1
5	standard	2.1
6	standard	1.7
7	standard	2.3
8	standard	2
9	standard	1.7
10	standard	2.2
11	new	2.5
12	new	2.1
13	new	2.4
14	new	2
15	new	2.6
16	new	2.3

Fig. 3.8 The available data in the GenStat server



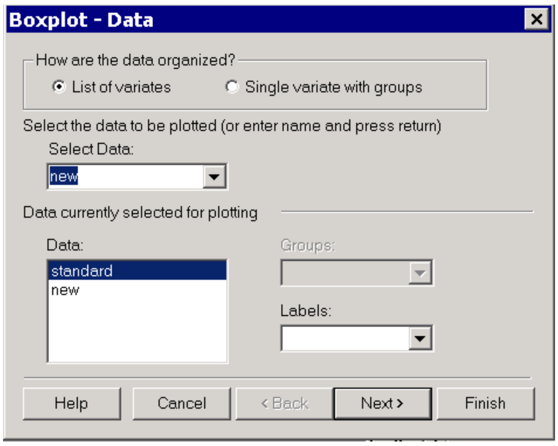
Name	Type	Attributes
Variety	Factor	Nvals = 16, Nlevels = 2
new	Variate	Nvals = 6
standard	Variate	Nvals = 10
yield	Variate	Nvals = 16

3.2 Visual data exploration.

3.2.1 Boxplots.

One way to present the data is to use a boxplot. It is always useful to explore data before carrying out any statistical test. This way you know what to expect and can discover anomalies. Use ***Graphics => Boxplot***. When the data are organized in two separate spreadsheets (like “Wheat variety new.gsh” and “Wheat variety standard.gsh”), complete the dialogue as shown in Fig. 3.9 and click ***[Finish]***. When the data are organised as a single variate with groups in one spreadsheet (in “wheat yield.gsh”), complete the dialogue as shown in Fig. 3.10. This gives the graphs in Fig. 3.11.

Fig. 3.9 Graphics => Boxplot when the data are in several variates



Boxplot - Data

How are the data organized?

● List of variates ○ Single variate with groups

Select the data to be plotted (or enter name and press return)

Select Data: new

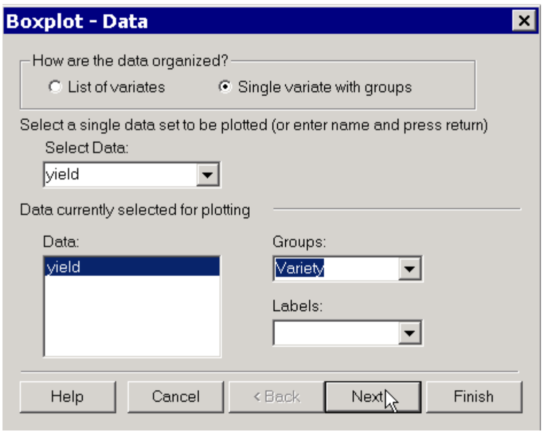
Data currently selected for plotting

Data:
standard
new

Groups:

Labels:

Fig. 3.10 Graphics => Boxplot when the data are organised as one variate with several groups



Boxplot - Data

How are the data organized?

○ List of variates ● Single variate with groups

Select a single data set to be plotted (or enter name and press return)

Select Data: yield

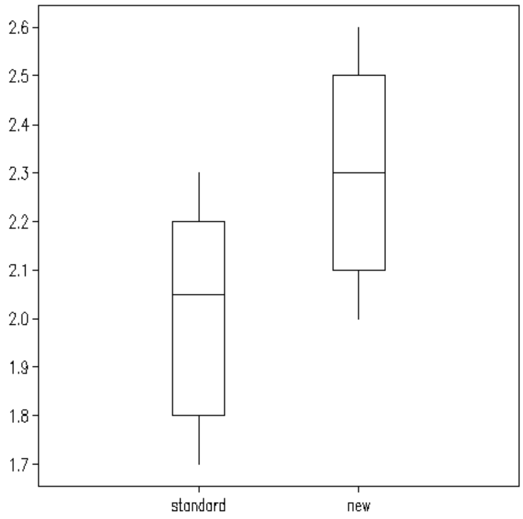
Data currently selected for plotting

Data: yield

Groups: Variety

Labels:

Fig. 3.11 The resulting boxplots



So, our visual impression is that the yield of the new variety is generally higher than the yield of the standard variety although there is some overlap. A formal statistical test has to confirm this, but first we go a bit deeper into boxplots.

3.2.2 Median and quartiles.

> A **boxplot** is the graphical representation of a 5-number summary of a dataset: minimum, Q1, median, Q3, maximum.

The middle value of data arranged in ascending order is called the **median**. When there is an even number of observations, the median is the average of the two middle values. Half the observations are smaller and half of the observations are larger than the median.

- $\tilde{x} = x_{(n+1)/2}$ (n = odd)
- $\tilde{x} = (x_{(n/2)} + x_{(n/2+1)})/2$ (n=even)

The median of the yield of the standard variety is (2.0 + 2.1)/2 = 2.05.

yield	1.7	1.7	1.8	1.9	**2.0**	**2.1**	2.1	2.2	2.2	2.3
rank	1	2	3	4	**5**	**6**	7	8	9	10

The value of the median is not influenced by extreme values nor does it change when the data are skew or bimodal.

Quartiles divide the data into quarters:

- 1st quartile = Q1 = 25 % of the observations are smaller, 75 % are bigger
- 2nd quartile = Q2 = median
- 3rd quartile = Q3 = 75 % of the observations are smaller, 25 % are bigger

Calculation of the quartiles[2]:

- Q1 = median of the group of observations below the median. Q1 of the standard wheat yield = 1.8
- Q3 = median of the group of observations above the median. Q3 of the standard wheat yield = 2.2

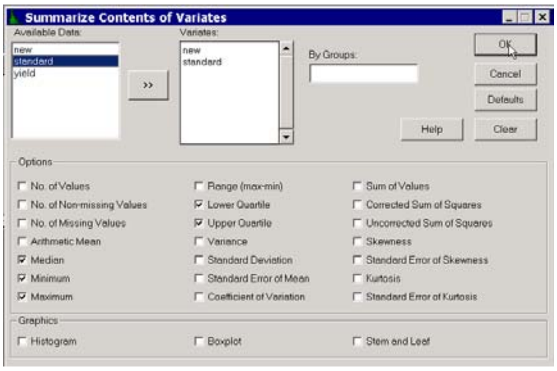
The difference between Q3 and Q1 is the **interquartile range (IQR)**. It is a measure of the spread of the data. It is not sensitive for extreme values. The IQR of the standard wheat yield = 0.4.

Median and quartile are special cases of **percentiles**. Generally, the pth percentile is a value whereby p % of the observations are lower than this value and (100 – p) % are higher. In GenStat Discovery Edition, percentiles are called **quantiles**.

There are several routes in GenStat of calculating median, quartiles and quantiles. One possibility is through the ***Stats => Summary Statistics => Summarize Contents of Variates*** menu. Fig. 3.12 shows the dialogue boxes when the 5-number summary is made for two variates while Fig. 3.13 shows this for one variate with groups. The results can be found in the Output Window.

[2] If the whole data set has an odd number of observations note that there are two ways of calculating the quartiles. GenStat excludes the median from the calculations of Q1 and Q3 while some other authors include the median in both calculations.

Fig. 3.12 Calculating the 5-number summary for data from two variates



Summarize Contents of Variates

Available Data: new, standard, yield

Variates: new, standard

By Groups:

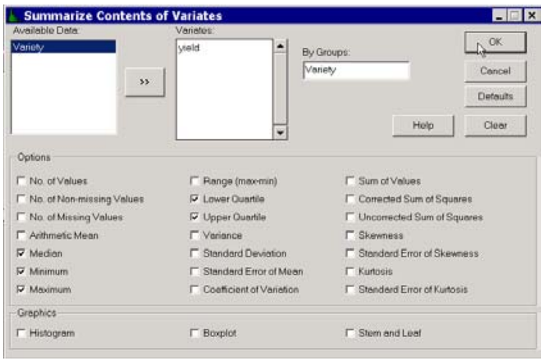
Options

- ☐ No. of Values
- ☐ No. of Non-missing Values
- ☐ No. of Missing Values
- ☐ Arithmetic Mean
- ☑ Median
- ☑ Minimum
- ☑ Maximum
- ☐ Range (max-min)
- ☑ Lower Quartile
- ☑ Upper Quartile
- ☐ Variance
- ☐ Standard Deviation
- ☐ Standard Error of Mean
- ☐ Coefficient of Variation
- ☐ Sum of Values
- ☐ Corrected Sum of Squares
- ☐ Uncorrected Sum of Squares
- ☐ Skewness
- ☐ Standard Error of Skewness
- ☐ Kurtosis
- ☐ Standard Error of Kurtosis

Graphics

- ☐ Histogram
- ☐ Boxplot
- ☐ Stem and Leaf

Fig. 3.13 Calculating the 5-number summary for data from one variate with two groups



Summarize Contents of Variates

Available Data: Variety

Variates: yield

By Groups: Variety

Options

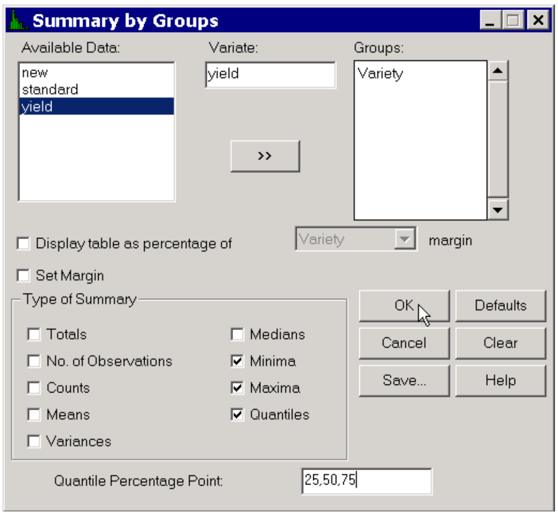
- ☐ No. of Values
- ☐ No. of Non-missing Values
- ☐ No. of Missing Values
- ☐ Arithmetic Mean
- ☑ Median
- ☑ Minimum
- ☑ Maximum
- ☐ Range (max-min)
- ☑ Lower Quartile
- ☑ Upper Quartile
- ☐ Variance
- ☐ Standard Deviation
- ☐ Standard Error of Mean
- ☐ Coefficient of Variation
- ☐ Sum of Values
- ☐ Corrected Sum of Squares
- ☐ Uncorrected Sum of Squares
- ☐ Skewness
- ☐ Standard Error of Skewness
- ☐ Kurtosis
- ☐ Standard Error of Kurtosis

Graphics

- ☐ Histogram
- ☐ Boxplot
- ☐ Stem and Leaf

For variates containing groups, also ***Stats => Summary Statistics => Summaries of Groups (Tabulation)*** can be chosen. Here you need to enter the quantile percentage points yourself. In the Quantile Percentage Point box of Fig. 3.14, enter the quantile percentage points for lower quartile, median and upper quartile (25, 50, 75). Clicking ***[OK]*** will give the results in the Output Window while clicking on the ***[Save]*** option (Fig. 3.15) will give you the possibility to save the summary statistics in several tables. The resulting tables are given in Fig. 3.16.

Fig. 3.14 Stats => Summary Statistics => Summaries of Groups (Tabulation)



Summary by Groups

Available Data: new, standard, yield

Variate: yield

Groups: Variety

☐ Display table as percentage of Variety margin

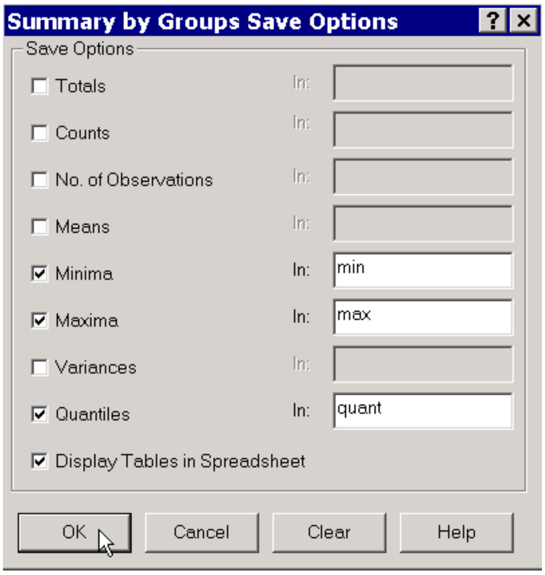
☐ Set Margin

Type of Summary

- ☐ Totals
- ☐ No. of Observations
- ☐ Counts
- ☐ Means
- ☐ Variances
- ☐ Medians
- ☑ Minima
- ☑ Maxima
- ☑ Quantiles

Quantile Percentage Point: 25,50,75

Fig. 3.15 Saving the summary statistics in several tables

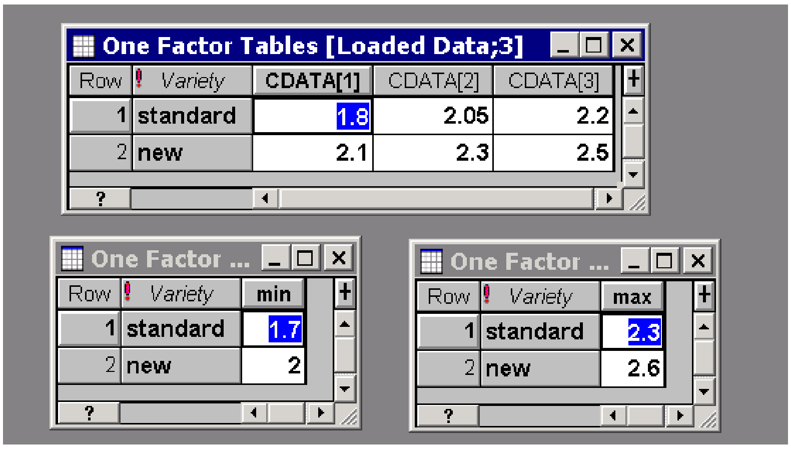


Summary by Groups Save Options

Save Options

- ☐ Totals In:
- ☐ Counts In:
- ☐ No. of Observations In:
- ☐ Means In:
- ☑ Minima In: min
- ☑ Maxima In: max
- ☐ Variances In:
- ☑ Quantiles In: quant
- ☑ Display Tables in Spreadsheet

Fig. 3.16 The resulting tables with minima, maxima, lower quartile, median and upper quartile per factor level



One Factor Tables [Loaded Data;3]

Row	Variety	CDATA[1]	CDATA[2]	CDATA[3]
1	standard	1.8	2.05	2.2
2	new	2.1	2.3	2.5

One Factor ...

Row	Variety	min
1	standard	1.7
2	new	2

One Factor ...

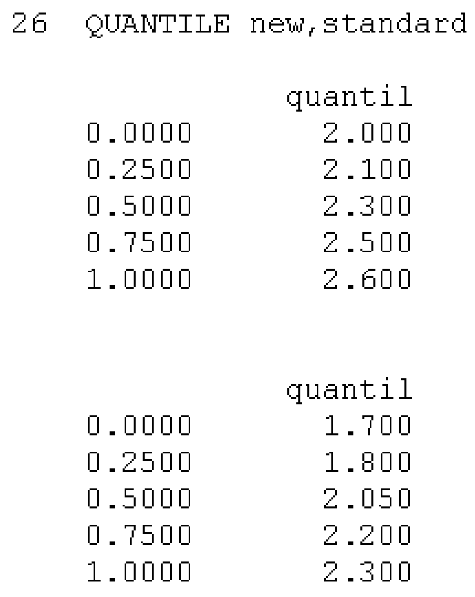
Row	Variety	max
1	standard	2.3
2	new	2.6

Finally, it is also possible to use commands as was introduced in chapter 2.4.2 (page 29). For instance, submitting the following line:

```
QUANTILE  standard,new
```

will return the 5-number summary of the variates new and standard in the Output Window (Fig. 3.17).

Fig. 3.17 The output of the QUANTILE procedure



```
 26  QUANTILE new,standard

                   quantil
       0.0000        2.000
       0.2500        2.100
       0.5000        2.300
       0.7500        2.500
       1.0000        2.600


                   quantil
       0.0000        1.700
       0.2500        1.800
       0.5000        2.050
       0.7500        2.200
       1.0000        2.300
```

3.2.3 The use of boxplots.

➔ **Comparing groups**

Boxplots are a useful tool **to compare groups of data**. In Fig. 3.11 for instance, it looked as if the yield of the new variety is higher than the yield of the standard variety.

There is however some overlap and remember the scale we are working with (minimum value is 1.7 tons/ha, maximum value is 2.5). A formal statistical test has to confirm the difference, but if this test shows completely different results to the graph, we know something is wrong.

➔ Outliers

Another use of boxplots is **to show outliers**. Go back to the spreadsheet and insert a value of 2.9 instead of 2.0 for the 8th value in the Standard group. Don't forget to press ***[Enter]*** after changing the value or the GenStat server will not be updated. The general shape of the graph is the same, but the odd value is indicated as deserving close scrutiny. There are now two ways to display the boxplot. Instead of using ***Graphics ⇒ Boxplot*** and immediately clicking ***[Finish]***, click ***[Next]***. You can choose now between two types: Box and Whisker and Schematic. The advantage of a schematic boxplot is that you can easily discover outliers.

Fig. 3.18 Box and Whisker plot

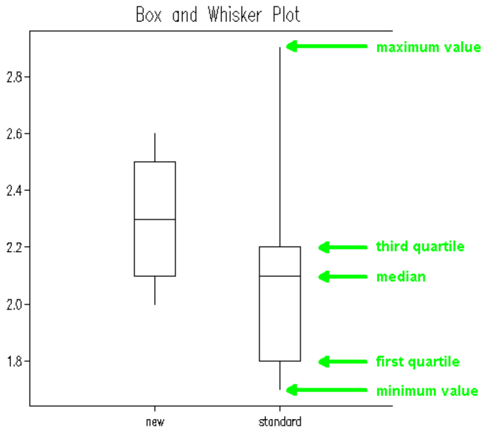
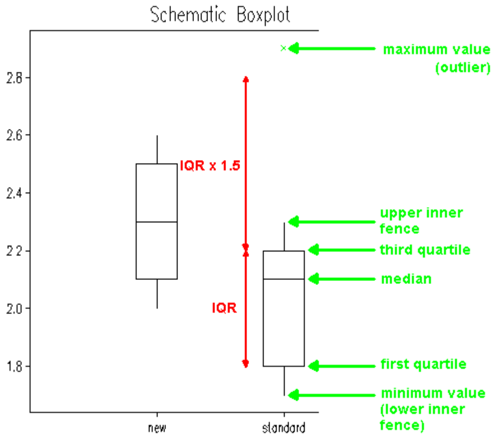


Fig. 3.19 Schematic boxplot with outlier marked



In a Box and Whisker boxplot, the ends of the whiskers mark the minimum and maximum values of the data set; in a schematic boxplot they mark the 'upper and lower inner fence'. The upper inner fence is defined as the maximum data value that is still smaller than the upper quartile plus 1.5 times the interquartile range; or the maximum value if this is less than upper quartile plus 1.5 times the IQR. The lower inner fence is defined similarly. Extreme values outside 1.5 and 3 times the interquartile range are by default marked as green crosses. More extreme values (more than 3 times the above mentioned range) are marked as red crosses.

➔ Shape of distribution

A boxplot gives an idea about **the shape of the distribution**, although you can also get this information from other plots (histograms and QQ-plot). Fig. 3.20 shows a boxplot from data from a normal distribution symmetric around 0. One observation can be considered as an outlier. Fig. 3.21 shows a histogram of the same data.

Fig. 3.20 Boxplot for normally distributed data (one outlier)

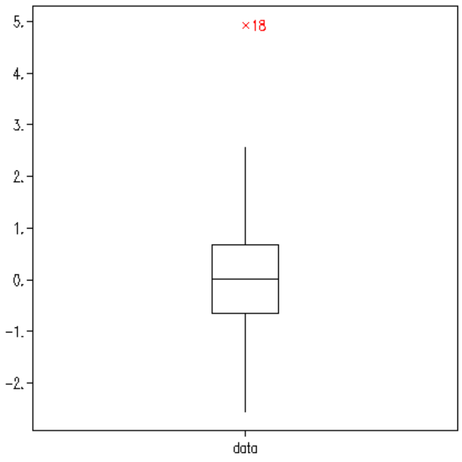
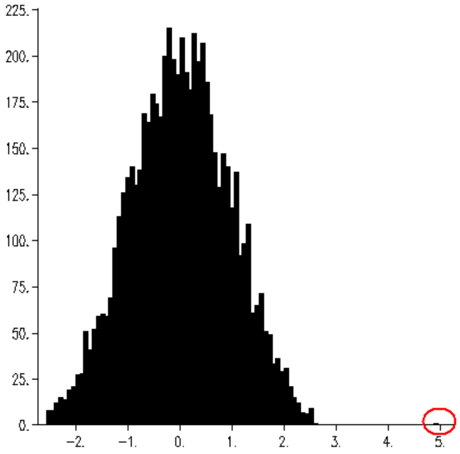


Fig. 3.21 The corresponding histogram for normally distributed data (one outlier)



In Fig. 3.22 we show an example based on a real world dataset of a skew distribution with a long tail of high outliers. 50 % of the observations have a value between 0 and 8 but the largest value is 106.

Fig. 3.22 Boxplot for skew data

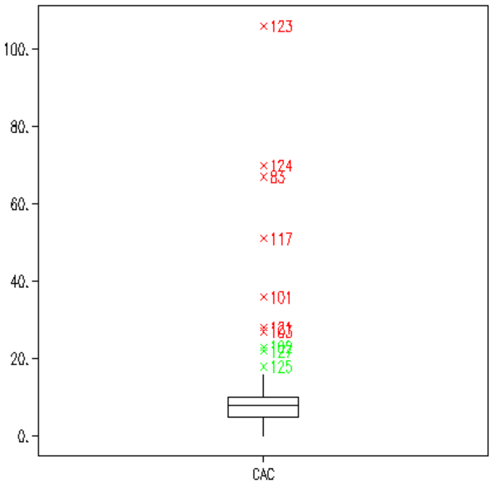
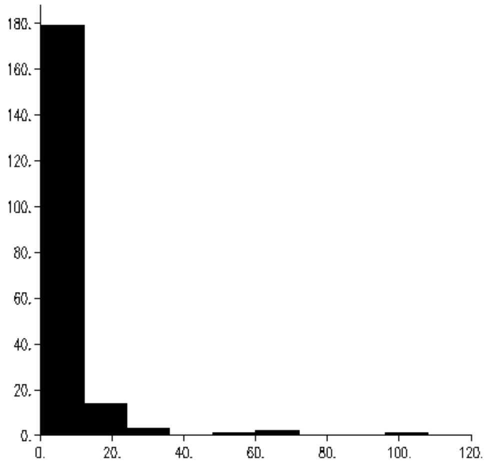


Fig. 3.23 The corresponding histogram for skew data



Before continuing, set the value of the standard wheat variety you changed back to 2.0.

3.3 Hypothesis testing.

Some of the examples we use are taken from “Confidence and Significance: Key Concepts of Inferential Statistics” from the Statistical Services Centre of The University of Reading, published in 2001. It can be downloaded from http://www.ssc.rdg.ac.uk/develop/dfid/booklets.html This booklet also contains more information on the statistical concepts.

3.3.1 Testing a hypothesis about the population mean.

The first example comes from this booklet. A researcher facilitates an on-farm trial to study the effect of using *Tephrosia vogelii* as a green manure for soil fertility restoration. She claims the use of the manure will increase pigeon pea yields measured as pod weight. In the trial pigeon peas are grown with and without the Tephrosia in two plots on each of eight smallholders’ fields and the values recorded are the differences in pod weights between plots (kg/plot):

3.0	3.6	5.4	-0.4	-0.8	4.2	4.8	3.2

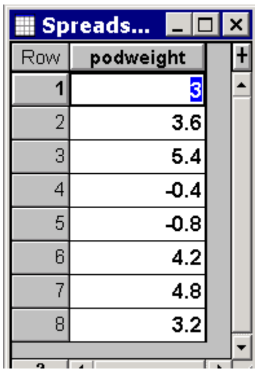
Our null hypothesis is that there is no difference in pod weights. We test this against the alternative hypothesis that there is a difference.

$$H_0 : \mu = 0 \, {}^{kg}\!/_{plot}$$

$$H_1 : \mu \neq 0 \, {}^{kg}\!/_{plot}$$

We first enter the differences in pod weight in a new spreadsheet and save it as *podweight.ghs* shown in Fig. 3.24 and we carry out some summary statistics. To carry out a t-test we need the mean and the standard error as given in Fig. 3.25.

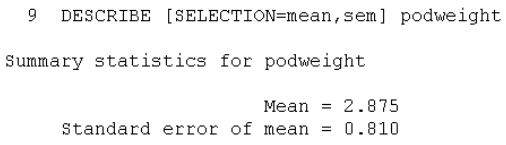
Fig. 3.24 The spreadsheet with differences in podweight



Spreads...

Row	podweight
1	3
2	3.6
3	5.4
4	-0.4
5	-0.8
6	4.2
7	4.8
8	3.2

Fig. 3.25 Some summary statistics in the Output Window



```
   9  DESCRIBE [SELECTION=mean,sem] podweight

Summary statistics for podweight

                          Mean = 2.875
      Standard error of mean = 0.810
```

The general formula for this one sample t-test is:

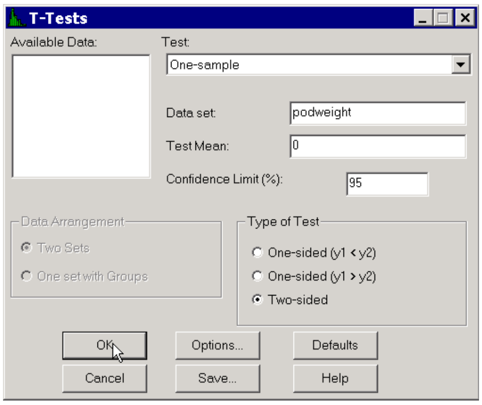
t = (estimate – hypothesised value)/ standard error of the estimate

In the example this becomes:

t = (2.875 – 0)/0.81 = 3.55 and we have to compare this with the t-distribution of 7 degrees of freedom.

In GenStat, choose ***Stats => Statistical Tests => T-Test*** and fill in the dialogue window as in Fig. 3.26

Fig. 3.26 The t-test dialogue box



Available Data:

Test: One-sample

Data set: podweight

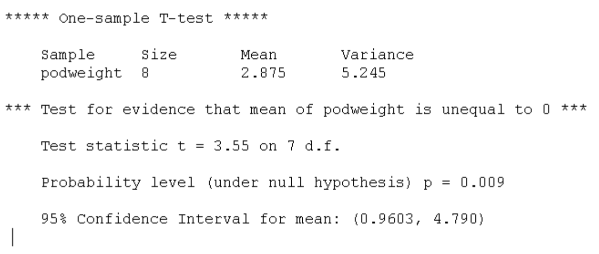
Test Mean: 0

Confidence Limit (%): 95

Data Arrangement
● Two Sets
○ One set with Groups

Type of Test
○ One-sided (y1 < y2)
○ One-sided (y1 > y2)
● Two-sided

Fig. 3.27 The results of a t-test in the output window



```
***** One-sample T-test *****

    Sample     Size       Mean       Variance
    podweight  8          2.875      5.245

*** Test for evidence that mean of podweight is unequal to 0 ***

    Test statistic t = 3.55 on 7 d.f.

    Probability level (under null hypothesis) p = 0.009

    95% Confidence Interval for mean: (0.9603, 4.790)
```

The results of the T-Test can be seen in the Output Window (Fig. 3.27). The p-value is 0.009, so if the null hypothesis is true (no differences in mean pod weight), then we have less than a 1 % chance to get the sample we have. This is not impossible. It is however so unlikely that we declare the result to be statistically significant and we reject the null hypothesis.

In the Output Window, we also find the 95 % confidence interval of the mean. This range is highly likely (95 % chance) to contain the true population mean. So, based on our sample, it is very likely that the average difference in pod weight between pigeon peas grown with Tephrosia and those not grown with Tephrosia is somewhere between 0.96 kg/plot and 4.79 kg/plot. The general formula for this 95 % confidence interval of the mean is

$$\bar{x} \pm t_{d.f.} \times s.e.\left(\bar{x}\right)$$

3.3.2 Comparison of samples.

We will use the example of the wheat yield again. Choose ***Run => Restart Session*** to clear all data from the GenStat server. Open the spreadsheet in the file ‘*Wheat yield.gsh*’.

In the boxplot in Fig. 3.11, it looked as if the yield of the new wheat variety is higher than that of the standard variety. We have reasons to believe this is true because the new variety was designed to produce higher yields. Our hypothesis is that the mean yield of the new variety is higher than the yield of the standard wheat variety. We will now do a formal statistical test. In this case, we will use ‘t’ test of two independent samples.

We rephrase our hypothesis as a set of null hypothesis and alternative hypothesis:

$$H_0 : \mu_{standard} - \mu_{new} = 0$$

$$H_1 : \mu_{standard} - \mu_{pooled} \neq 0$$

For this t-test, the general formula is:

t = (estimated mean of first sample – estimated mean of second sample)/(standard error of difference of the means)

The calculations necessary to perform the test depend on two assumptions:

- both samples come from normally distributed populations
- both populations have the same variance

Because of this last assumption, we can combine the two sample variances to give a better estimate of the variance in the two populations. This **pooled variance** is calculated as:

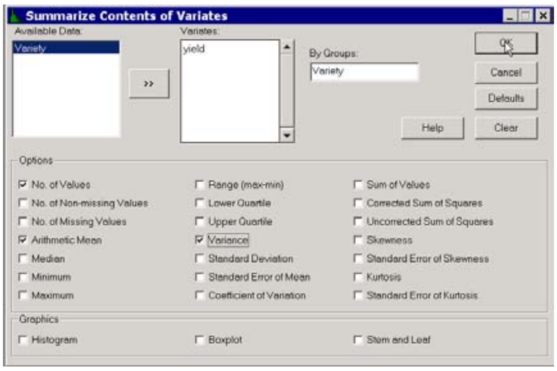
$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

The standard error of the differences of means is then:

$$s.e.d. = \sqrt{\frac{s^2_{pooled}}{n_1} + \frac{s^2_{pooled}}{n_2}}$$

It is possible to get the necessary summary statistics (Fig. 3.28) and calculate the pooled variance (Fig. 3.29) of 0.0502143. This can then in turn be used to calculate a t-value of -2.59253 (not shown). This t-value has to be compared with the t-distribution with 14 degrees of freedom ($n_1 + n_2 – 2 = 6 + 10 – 2$) that can be found in most statistical textbooks.

Fig. 3.28 The necessary summary statistics



Summarize Contents of Variates

Available Data: Variety

Variates: yield

By Groups: Variety

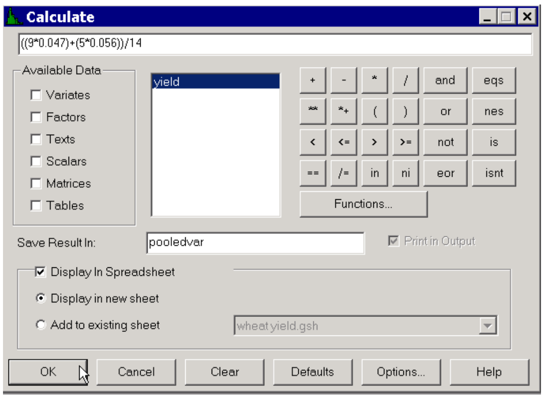
Options

☑ No. of Values
☐ No. of Non-missing Values
☐ No. of Missing Values
☑ Arithmetic Mean
☐ Median
☐ Minimum
☐ Maximum
☐ Range (max-min)
☐ Lower Quartile
☐ Upper Quartile
☑ Variance
☐ Standard Deviation
☐ Standard Error of Mean
☐ Coefficient of Variation
☐ Sum of Values
☐ Corrected Sum of Squares
☐ Uncorrected Sum of Squares
☐ Skewness
☐ Standard Error of Skewness
☐ Kurtosis
☐ Standard Error of Kurtosis

Graphics

☐ Histogram
☐ Boxplot
☐ Stem and Leaf

Fig. 3.29 Calculating the pooled variance



Calculate

((9*0.047)+(5*0.056))/14

Available Data

☐ Variates
☐ Factors
☐ Texts
☐ Scalars
☐ Matrices
☐ Tables

yield

Save Result In: pooledvar

☑ Print in Output

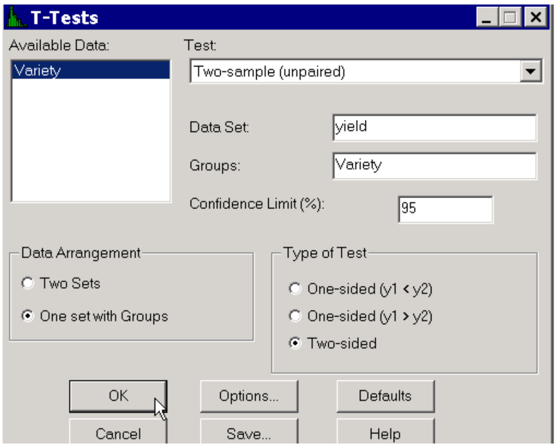
☑ Display In Spreadsheet

● Display in new sheet

○ Add to existing sheet: wheat yield.gsh

This was the hard way that could be of use when teaching statistics but it is of course easier to let the software do all the work. Choose ***Stats => Statistical Test => T-test***, select in the Test box the Two-sample (unpaired) test and make sure you indicate the data consist of one set with groups as in Fig. 3.30. If you prefer to work with both “*Wheat variety new.gsh*” and “*Wheat variety standard.gsh*”, proceed as in Fig. 3.31.

Fig. 3.30 Comparing two samples when there is one variate with two groups



T-Tests

Available Data: Variety

Test: Two-sample (unpaired)

Data Set: yield

Groups: Variety

Confidence Limit (%): 95

Data Arrangement

○ Two Sets

● One set with Groups

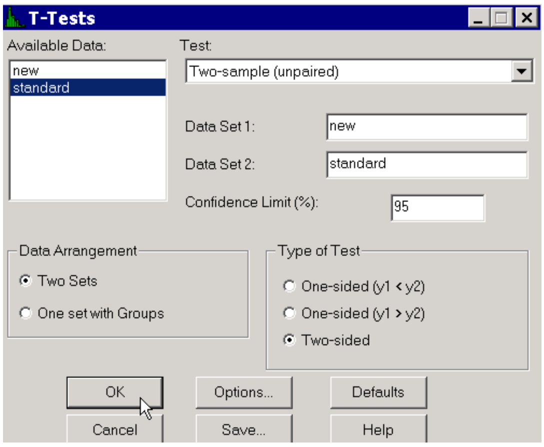
Type of Test

○ One-sided (y1 < y2)

○ One-sided (y1 > y2)

● Two-sided

Fig. 3.31 Comparing two samples when there are two variates



T-Tests

Available Data: new, standard

Test: Two-sample (unpaired)

Data Set 1: new

Data Set 2: standard

Confidence Limit (%): 95

Data Arrangement

● Two Sets

○ One set with Groups

Type of Test

○ One-sided (y1 < y2)

○ One-sided (y1 > y2)

● Two-sided

The results of the test are given in the Output Window.

```
***** Two-sample T-test *****


    Sample      Size       Mean       Variance
    standard    10         2.000      0.04667
    new         6          2.300      0.05600


*** Test for equality of sample variances ***


    Test statistic F = 1.20 on 5 and 9 d.f.


    Probability level (under null hypothesis of equal variances) =
0.76


*** Test for evidence that mean of yield with Variety = standard
is unequal to mean with Variety = new ***


    Test statistic t = -2.60 on 14 d.f.


    Probability level (under null hypothesis) p = 0.021


    95% Confidence Interval for difference in means: (-0.5477, -
0.05234)
```

If the null hypothesis is true (both population means are equal), then we only have a chance of about 2 % to find the samples we found (p-value is 0.021). So we can reject the null hypothesis and decide that there is a statistically significant difference between the two sample means.

What else do we find in the Output Window ? First we find some summary statistics, next we see the results of an F-test, then the t-test and finally the 95 % confidence interval for the difference in means.

By default, GenStat gives an F-test for equality of sample variances, because this was one of the assumptions we used to perform the t-test of two independent samples. However, this F-test only performs well if the distributions of the populations are close to the normal distribution.

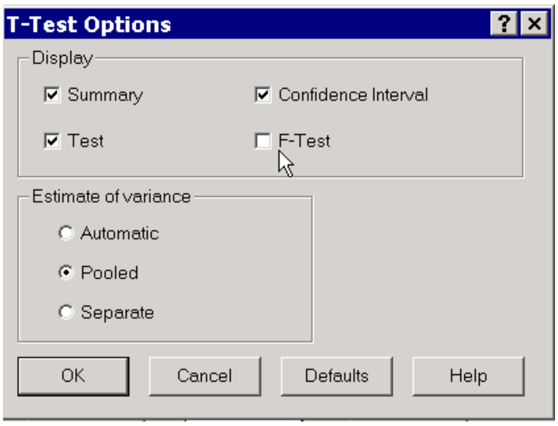
The general formula for the 95% confidence interval of differences in means is given by:

$$\overline{x}_1 - \overline{x}_2 \pm t_{n_1+n_2-2} \times s.e.d.$$

So, based on our samples, it is very likely that the standard wheat variety will produce on average 0.05 to 0.55 tonnes/ha less than the new wheat variety.

If necessary, some of the results in the Output Window (for instance this F-test) can be suppressed by changing the options of the t-test. After choosing ***Stats=>Statistical Tests=>T-test***, click on the ***[Options]*** button (Fig. 3.32).

Fig. 3.32 Comparing two samples when there is one variate with two groups



Display

☑ Summary ☑ Confidence Interval

☑ Test □ F-Test

Estimate of variance

○ Automatic

● Pooled

○ Separate

3.3.3 Paired t-test.

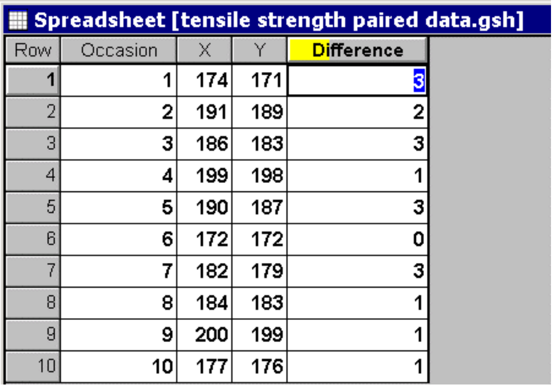
In the above example, we compared two independent groups. In this example we will perform the test on paired data. The example comes from Confidence and Significance: Key Concepts of Inferential Statistics (Statistical Services Centre, University of Reading, 2001) page 17 (data on page 14). The x and y values represent the tensile strength of rubber samples collected from two plantations (x and y) on 10 occasions. The aim was to see whether the two plantations differed in the quality of their rubber samples.

Occasion	X	Y
1	174	171
2	191	189
3	186	183
4	199	198
5	190	187
6	172	172
7	182	179
8	184	183
9	200	199
10	177	176

Performing a paired test means that the plantation to plantation variability is removed from the analysis, so we compare the differences in tensile strength at each occasion.

In GenStat, choose ***Run => Restart Session*** and create a new spreadsheet where you enter the above-mentioned data. To show two ways of doing the test, first calculate the difference in strength between the X and Y plantation on each occasion in a new variable called "Difference". Save the file as '*tensile strength paired data.gsh*' (Fig. 3.33).

Fig. 3.33 tensile strength paired data.gsh

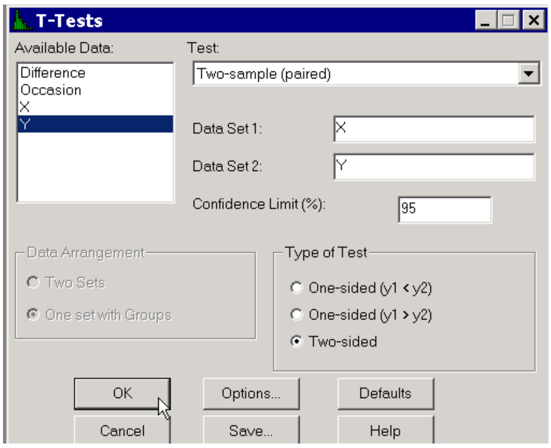


Spreadsheet [tensile strength paired data.gsh]

Row	Occasion	X	Y	Difference
1	1	174	171	3
2	2	191	189	2
3	3	186	183	3
4	4	199	198	1
5	5	190	187	3
6	6	172	172	0
7	7	182	179	3
8	8	184	183	1
9	9	200	199	1
10	10	177	176	1

The first way of doing the test is to select a two-sample (paired test) and compare X with Y (Fig. 3.34). Or you can select a one-sample test and compare the differences with a mean of zero (Fig. 3.35).

Fig. 3.34 Paired two sample t-test



T-Tests

Available Data: Difference, Occasion, X, Y

Test: Two-sample (paired)

Data Set 1: X

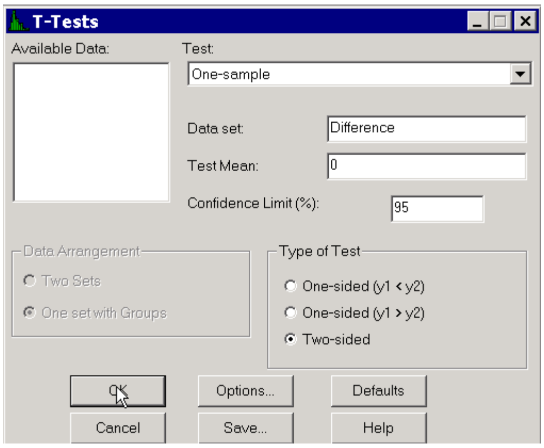
Data Set 2: Y

Confidence Limit (%): 95

Data Arrangement: ○ Two Sets ● One set with Groups

Type of Test: ○ One-sided (y1 < y2) ○ One-sided (y1 > y2) ● Two-sided

Fig. 3.35 The one-sample approach for a paired t-test



T-Tests

Available Data:

Test: One-sample

Data set: Difference

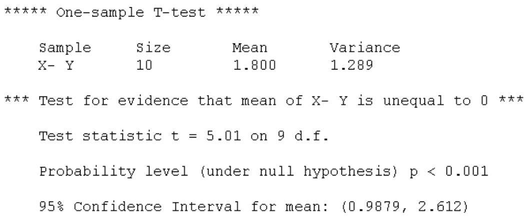
Test Mean: 0

Confidence Limit (%): 95

Data Arrangement: ○ Two Sets ● One set with Groups

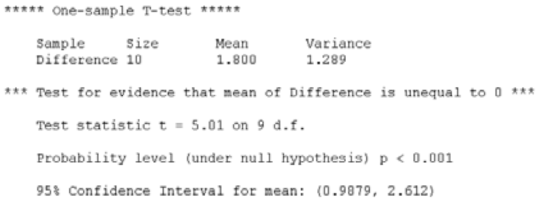
Type of Test: ○ One-sided (y1 < y2) ○ One-sided (y1 > y2) ● Two-sided

Fig. 3.36 Output of a paired two sample t-test



```
***** One-sample T-test *****

    Sample      Size        Mean        Variance
    X- Y        10          1.800       1.289

*** Test for evidence that mean of X- Y is unequal to 0 ***

    Test statistic t = 5.01 on 9 d.f.

    Probability level (under null hypothesis) p < 0.001

    95% Confidence Interval for mean: (0.9879, 2.612)
```

Fig. 3.37 Output of the one-sample approach for a paired t-test



```
***** One-sample T-test *****

    Sample      Size        Mean        Variance
    Difference  10          1.800       1.289

*** Test for evidence that mean of Difference is unequal to 0 ***

    Test statistic t = 5.01 on 9 d.f.

    Probability level (under null hypothesis) p < 0.001

    95% Confidence Interval for mean: (0.9879, 2.612)
```

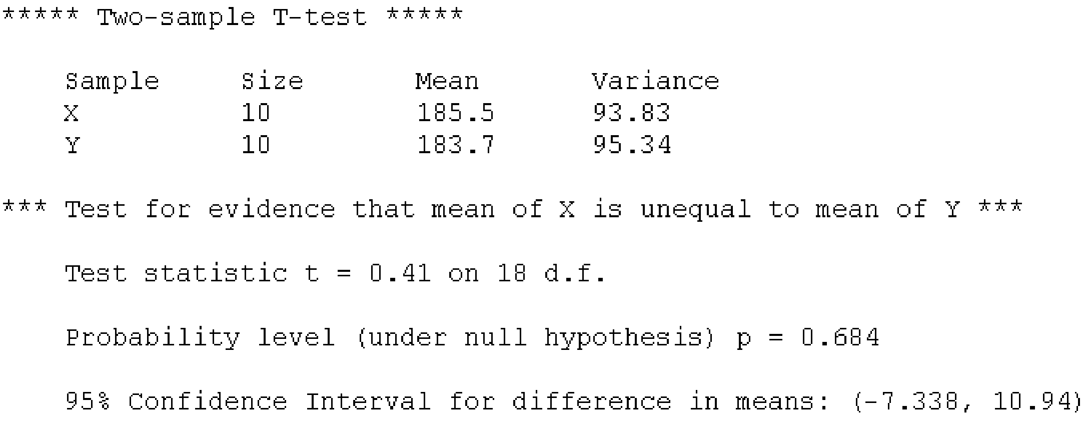
The Output Window gives exactly the same results in both cases (Fig. 3.36 and Fig. 3.37). No wonder, because by choosing a paired t-test we indicated we want to ignore the plantation to plantation variability. So in both cases we tested that the mean of the pair wise differences is equal to zero.

The t-statistic was calculated using the mean difference and the standard error of the estimate as seen in chapter 3.3.1:

$$t = (1.8)/\sqrt{1.289/10} = 5.013$$

By comparing matched pairs we improved the precision of the analysis. If we had chosen to perform a t-test of two independent samples, the small but systematic differences between the pairs would not have been detected. We would have calculated a pooled variance from the rather large variances of X and Y. This would have lead to a non-significant t-value of 0.41 (Fig. 3.38) and we would have mistakenly concluded that there are no differences in tensile strengths between the two plantations.

Fig. 3.38 Using the wrong approach leads to wrong results although nothing is technically wrong with the calculations.



```
***** Two-sample T-test *****

    Sample      Size        Mean        Variance
    X           10          185.5       93.83
    Y           10          183.7       95.34

*** Test for evidence that mean of X is unequal to mean of Y ***

    Test statistic t = 0.41 on 18 d.f.

    Probability level (under null hypothesis) p = 0.684

    95% Confidence Interval for difference in means: (-7.338, 10.94)
```

This paired structure can be compared to the concept of blocking in experiments and stratification in surveys.

3.3.4 A non-parametric example.

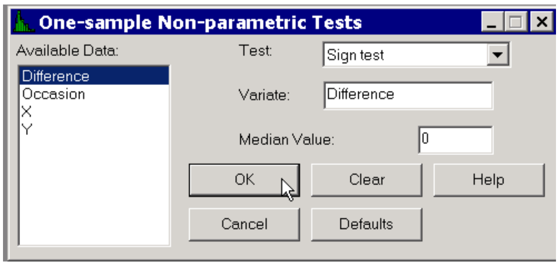
All t-tests and generally much statistical analysis are based on the assumption that data come from a normal distribution. Sometimes this is not the case, for instance:

- the distribution is very skew because one or some measurements are much larger that the usual range and are not a measurement error.
- measurements are not on the ratio scale but on the ordinal scale. For instance different farmers may assign scores between 1 and 10 about preferences of using different tree species on their farm. Some farmers will avoid extreme scores, others will use them.

In such cases we may choose to use non-parametric methods.

Let's imagine that the differences in tensile strength (Fig. 3.33) consist of such data. A possible approach now is to use the **sign test**. Under a null-hypothesis of no difference between the two samples, about half of the differences would be positive and about half would be negative, so the median would be zero. In the example, 9 differences are positive, one difference is zero and zero differences are negative. We do not go into details about the calculations but show how to perform the test in GenStat. Choose ***Stats => Statistical Tests => One-sample non-parametric tests***. The variate to be tested is "*Difference*" and by default, GenStat tests against a median value of zero (Fig. 3.39).

Fig. 3.39 Sign test



One-sample Non-parametric Tests

Available Data: Difference, Occasion, X, Y

Test: Sign test

Variate: Difference

Median Value: 0

```
***** One-sample Sign Test *****


      Variate        Size      Median
   Difference           9       1.500


 Test if median equals 0


                 Test statistic:      9
          Effective sample size:      9
    Two-sided probability level:  0.004
```

In this example it is very obvious that the null-hypothesis of no difference is rejected (p=0.04).

3.4 A fast and simple regression.

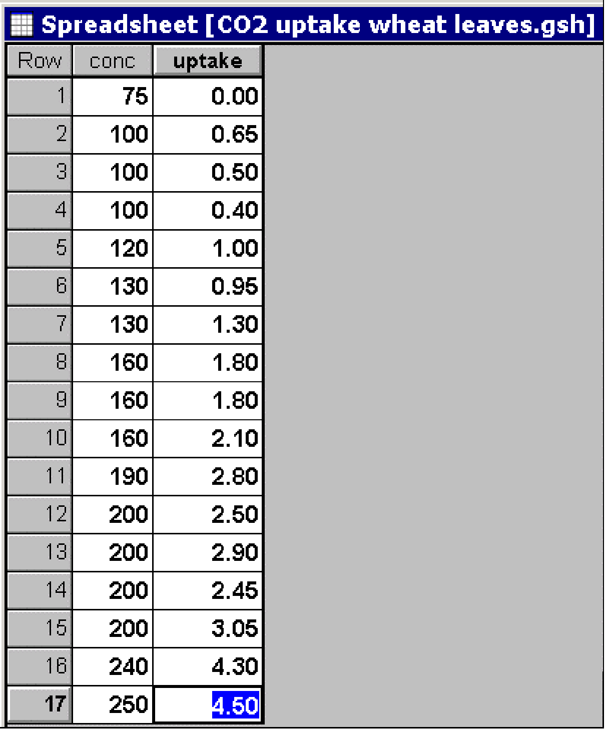
We will now introduce some key elements of data analysis using GenStat, by means of simple regression. This example is taken from page 193 of Mead, Curnow and Hasted. In this chapter we only show how to perform a linear regression in GenStat and what kind of options are available.

Start a new GenStat session (see chapter 2.4.3), create a spreadsheet with 2 columns (*conc* and *uptake*) and 17 rows and enter the data from Fig. 3.40. Add extra description to the two columns: *conc* are various CO_2 concentrations passed over wheat leaves at a temperature of 35 °C and *uptake* is the amount of CO_2 that is taken up by those leaves. Format the *uptake* column so it shows two decimals. Save the spreadsheet as '*CO2 uptake wheat leaves.gsh*'. See chapter 2.2 if you need some help with this.

Fig. 3.40 Simple regression data

CO_2 concentration	Uptake (cm3/dm2/hour)
conc	uptake
75	0.00
100	0.65
100	0.50
100	0.40
120	1.00
130	0.95
130	1.30
160	1.80
160	1.80
160	2.10
190	2.80
200	2.50
200	2.90
200	2.45
200	3.05
240	4.30
250	4.50

Fig. 3.41 The same data in a GenStat spreadsheet

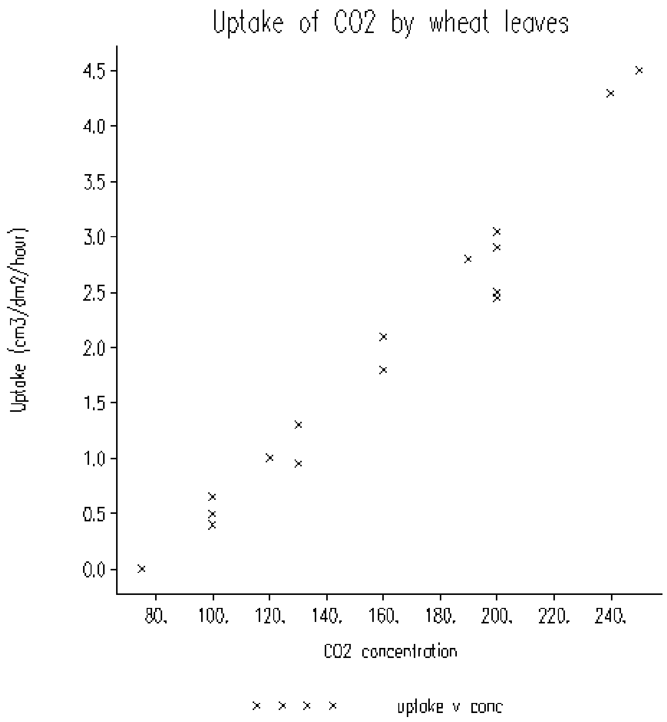


Spreadsheet [CO2 uptake wheat leaves.gsh]

Row	conc	uptake
1	75	0.00
2	100	0.65
3	100	0.50
4	100	0.40
5	120	1.00
6	130	0.95
7	130	1.30
8	160	1.80
9	160	1.80
10	160	2.10
11	190	2.80
12	200	2.50
13	200	2.90
14	200	2.45
15	200	3.05
16	240	4.30
17	250	4.50

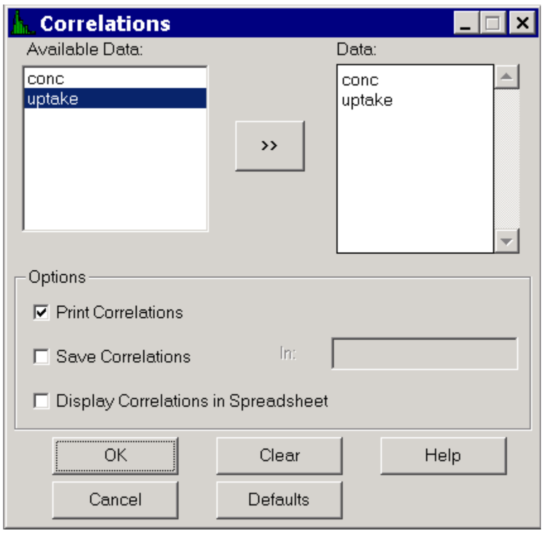
Before starting a formal analysis we first look at the data in an exploratory way. Check the summary statistics for each of the two columns (see chapter 2.3.1) and draw a graph. Draw a point plot as in Fig. 3.42 to see if there is a linear relationship.

Fig. 3.42 A point plot of the regression data



Choose ***Stats* ⇒ *Summary Statistics* ⇒ *Correlations*** and complete the dialogue as shown in Fig. 3.43 to give the correlation between *uptake and conc*. Indicate you want to see the correlations in a spreadsheet.

Fig. 3.43 Correlations dialogue


Correlations

Available Data: conc, uptake

Data: conc, uptake

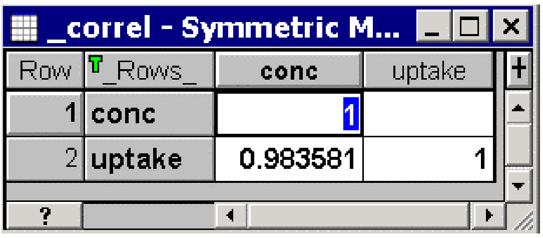
Options

☑ Print Correlations

□ Save Correlations In:

□ Display Correlations in Spreadsheet

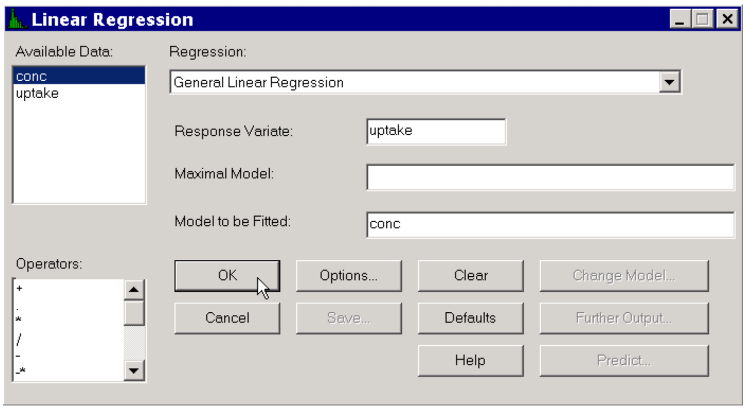
Fig. 3.44 Results in a new spreadsheet


_correl - Symmetric M...

Row	_Rows_	conc	uptake
1	conc	1	
2	uptake	0.983581	1

So, we see a linear pattern and a high positive correlation between the CO_2 concentration and the CO_2 uptake. We have therefore fit a straight-line model to the data. Choose ***Stats* ⇒ *Regression Analysis* ⇒ *Linear***. Choose or Simple Linear Regression or General Linear Regression in the regression box and click ***[OK]***. Once you have done this, the results of the regression can be seen in the Output Window and the buttons on the linear regression menu that were dimmed in Fig. 3.45 have become active.

Fig. 3.45 Linear regression dialogue



Linear Regression

Available Data: conc, uptake

Regression: General Linear Regression

Response Variate: uptake

Maximal Model:

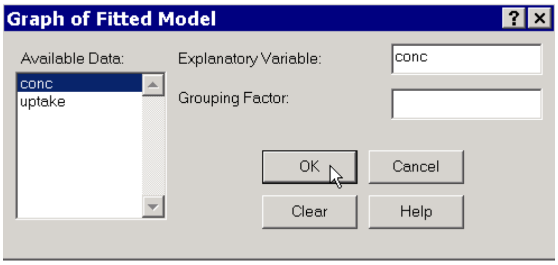
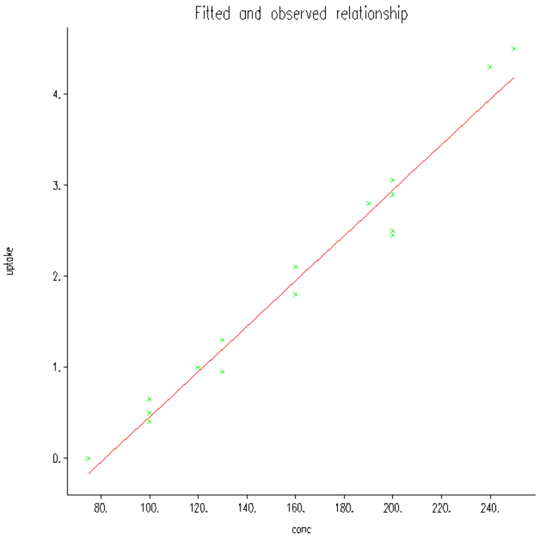
Model to be Fitted: conc

Operators: + . * / - -*

At the bottom of the Output Window we can see the estimate of parameters in the fitted equation:

uptake = -2.043 + 0.02494 * conc

Click on ***[Further Output]*** in the dialogue in Fig. 3.45, next on ***[Fitted Model]*** and finally give the explanatory variable, see Fig. 3.46 to produce a graph with the original observations and the fitted regression line.

Fig. 3.46 Further output from the regression model	*Fig. 3.47 The resulting graph*
	

Graph of Fitted Model

Available Data: conc, uptake

Explanatory Variable: conc

Grouping Factor:

This example shows it is easy to “do statistics” once you have become familiar with the use of dialogues in GenStat.

4 Review of chapters 2–3.

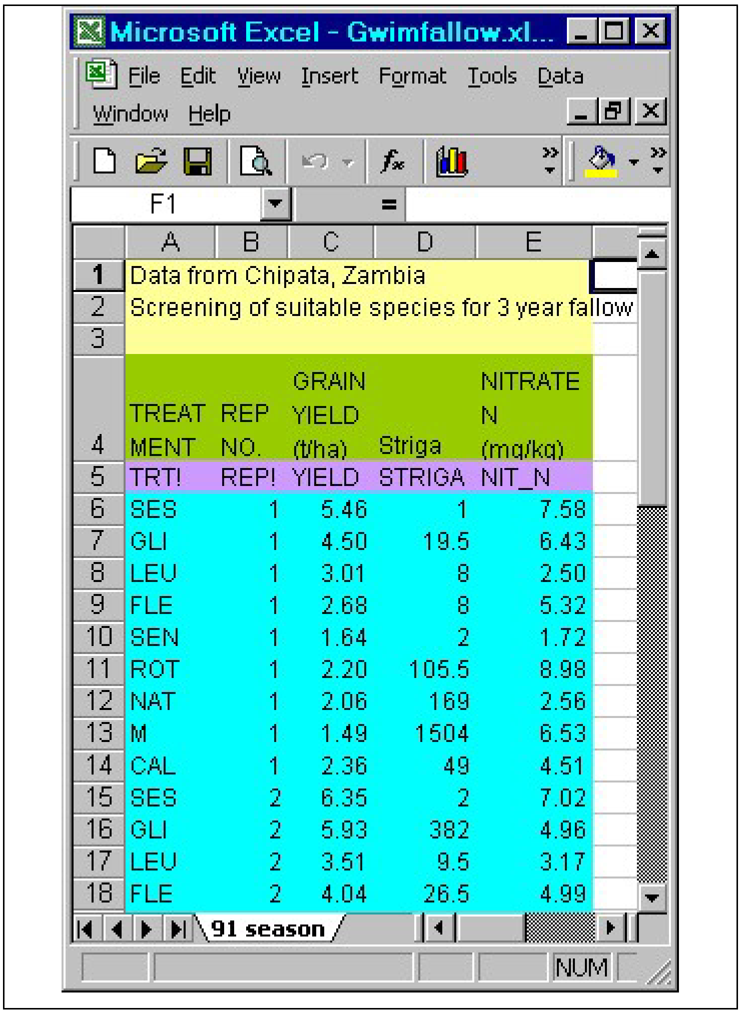
Here we review some of the tasks you have undertaken in the previous chapters. **Could you?**

Task	**Hint**
Open a set of data you entered earlier in Excel, for example the file *“Prunus africana height and dbh Mabira Uganda.xls”*?	See page 10
Enter a new set of data which has 3 columns and 6 rows?	See page 7
Import a named range from an Excel worksheet.	See page 12
Derive a new column containing the square of the values in an existing column?	See page 17
Append a GenStat spreadsheet to another?	See page 35
Carry out a two-sample unpaired **t-test?**	See page 45
Find the names and lengths of all the columns of data.	See page 28
Explain why a boxplot is often a useful summary of a set of data and also to compare different sets?	See page 42, look in a statistics book, or ask someone.
Give a **line plot?**	Look at the second option in the menu shown in Fig. 2.27 on page 16
Carry out a simple linear regression?	See page 52

Task	**Hint**
Summarise a column of data?	See page 15 and page 26
Explain **how GenStat “works”?**	See chapter 2.4 on page 28
Explain what is meant by **a factor column?**	See page 21
Leave GenStat? (If not, then keep practising!)	

5 Challenge 1

The file called “Fallow species trial.xls” contains data from a field experiment in which soil nitrate was measured at the start of the season in plots that had been under various fallows (coded in TRT). Maize yield was measured in each plot at the end of the season, as well as the level of Striga (a parasitic weed) infestation. Find the average grain yield for each type of fallow. Produce a graph that shows the relationship between maize yield and preseason soil nitrate for each type of fallow. Check whether there is an obvious relationship between maize yield and Striga, and whether the plot looks clearer if square root (Striga) is considered.


Microsoft Excel - Gwimfallow.xl...
File Edit View Insert Format Tools Data
Window Help
F1
=
91 season
NUM

	A	B	C	D	E
1	Data from Chipata, Zambia				
2	Screening of suitable species for 3 year fallow				
3					
4	TREATMENT	REP NO.	GRAIN YIELD (t/ha)	Striga	NITRATE N (mg/kg)
5	TRT!	REP!	YIELD	STRIGA	NIT_N
6	SES	1	5.46	1	7.58
7	GLI	1	4.50	19.5	6.43
8	LEU	1	3.01	8	2.50
9	FLE	1	2.68	8	5.32
10	SEN	1	1.64	2	1.72
11	ROT	1	2.20	105.5	8.98
12	NAT	1	2.06	169	2.56
13	M	1	1.49	1504	6.53
14	CAL	1	2.36	49	4.51
15	SES	2	6.35	2	7.02
16	GLI	2	5.93	382	4.96
17	LEU	2	3.51	9.5	3.17
18	FLE	2	4.04	26.5	4.99

6 Before starting an Analysis of Variance

GenStat has comprehensive facilities for the analysis of designed experiments. In this chapter we look first at the way the data are set-up for this type of analysis. This extends the discussion of factors that was introduced in chapter 2. We then consider examples of a completely randomised, randomised block and split plot experiment. The important concept of factorial treatment structure is also described.

6.1 Factors and data organization

6.1.1 In a GenStat spreadsheet.

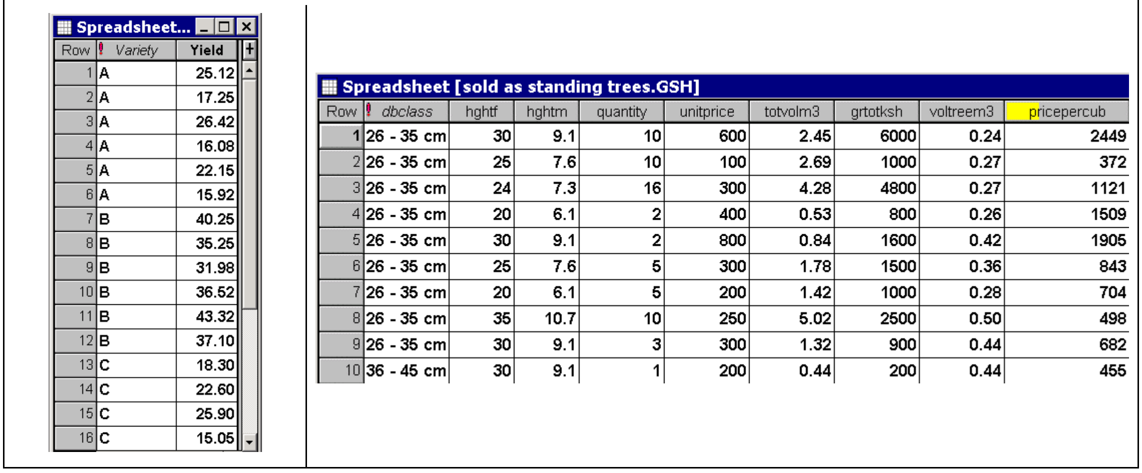
In a textbook or course on statistics, a dataset containing measurements on the yield of 4 melon varieties may be given as in Fig. 6.1 (see Mead, Curnow and Hasted, 2003. p. 58):

Fig. 6.1 Common data layout in textbooks

Variety	A	B	C	D
Yields	25.12	40.25	18.30	28.55
	17.25	35.25	22.60	28.05
	26.42	31.98	25.90	33.20
	16.08	36.52	15.05	31.68
	22.15	43.32	11.42	30.32
	15.92	37.10	23.68	27.58

This tabular layout is not, however, best suited, or indeed, acceptable to most statistical packages. Instead, as shown in Fig. 6.2, the measurements are entered in columns of length equal to the total number of units. Together with these data, other columns are entered that describe the experimental treatment, etc. These are the **factors** that were introduced earlier (chapter 2.3.3 page 21). Common examples are the unit number, the block from which the unit comes, or the fertiliser amount applied to the plot. In most experiments, there is more than one measurement. The way of entering data with each measurement in a single column is also well suited to such situations. An example is shown in Fig. 6.2.

Fig. 6.2 Data layout for use with statistical packages



Row	! *Variety*	Yield
1	A	25.12
2	A	17.25
3	A	26.42
4	A	16.08
5	A	22.15
6	A	15.92
7	B	40.25
8	B	35.25
9	B	31.98
10	B	36.52
11	B	43.32
12	B	37.10
13	C	18.30
14	C	22.60
15	C	25.90
16	C	15.05

Row	! *dbclass*	hghtf	hghtm	quantity	unitprice	totvolm3	grtotksh	voltreem3	pricepercub
1	26 - 35 cm	30	9.1	10	600	2.45	6000	0.24	2449
2	26 - 35 cm	25	7.6	10	100	2.69	1000	0.27	372
3	26 - 35 cm	24	7.3	16	300	4.28	4800	0.27	1121
4	26 - 35 cm	20	6.1	2	400	0.53	800	0.26	1509
5	26 - 35 cm	30	9.1	2	800	0.84	1600	0.42	1905
6	26 - 35 cm	25	7.6	5	300	1.78	1500	0.36	843
7	26 - 35 cm	20	6.1	5	200	1.42	1000	0.28	704
8	26 - 35 cm	35	10.7	10	250	5.02	2500	0.50	498
9	26 - 35 cm	30	9.1	3	300	1.32	900	0.44	682
10	36 - 45 cm	30	9.1	1	200	0.44	200	0.44	455

> Notice that in the example the name of the columns '*Variety*' and '*dbclass'* are in italics. In a GenStat spreadsheet, a name in italics preceded by an exclamation mark (!) indicates that the column is a factor.

We will now create some spreadsheets that will be used in chapter 8.

6.1.1.1 Melon yields.

Create a spreadsheet with 2 columns and 24 rows. The first column, called *Variety*, is a factor with 4 levels ("A", "B", "C" and "D"). The second column *Yield* is a variate. Format *Yield* to show 2 decimals. Save the spreadsheet as "*Melon yield.gsh*" and choose ***Run => Restart Session*** to clear all data from memory.

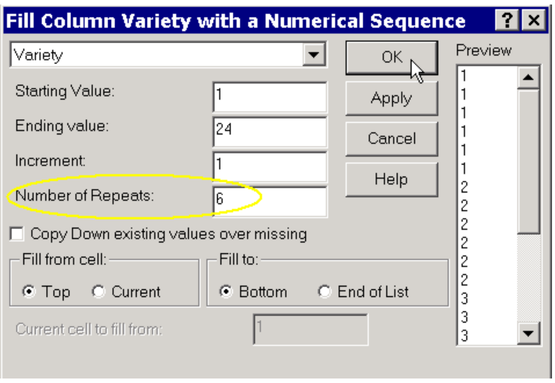
> Normally data are entered in the randomised order that corresponds to the recording in the field book and a column is included that gives the plot number. Here however we have entered the data in the order shown in the textbook.

All this was shown in chapter 2.3.3. In that chapter we also saw the difference between entering ordinals or labels. We show now some alternative ways of creating a spreadsheet, some are more suitable when entering ordinals, others when entering labels. Skip the rest of this section if you're happy to know only one way of entering data into GenStat.

➔ **Fill**

This is the easiest option when entering data as ordinals. In the example we could fill a column containing a variate: 6 times 1, 6 times 2, 6 times 3 and 6 times 4. Choose ***Spread => Calculate => Fill*** and indicate you want 6 repeats (Fig. 6.3). In the preview window you are able to see how the figures will appear in the column. Next choose ***Spread => Column => Convert***, convert the column into a factor type and with the ***Spread => Factor => Edit labels*** option you can change the 4 figures into 4 letters (Fig. 6.4). Press the ***[Enter]*** key after typing each label.

Fig. 6.3 Indicating the number of repeats



Fill Column Variety with a Numerical Sequence

Variety

Starting Value: 1

Ending value: 24

Increment: 1

Number of Repeats: 6

□ Copy Down existing values over missing

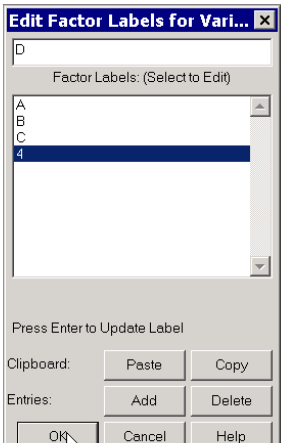
Fill from cell: ● Top ○ Current

Fill to: ● Bottom ○ End of List

Current cell to fill from: 1

Preview: 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3

Fig. 6.4 Changing the Factor labels



Edit Factor Labels for Vari...

D

Factor Labels: (Select to Edit)

A
B
C
4

Press Enter to Update Label

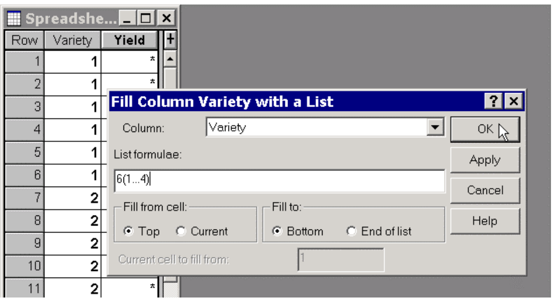
> If you already have a factor, and you want to give it labels or to modify the existing labels, click anywhere in the column, choose ***Spread*** ⇒ ***Factor*** ⇒ ***Edit Labels***.

➔ **List fill.**

This option is slightly more difficult. Consider it as a good practice for when you want to learn the GenStat command language. Otherwise, the Fill option is easier. ***Spread => Calculate => List Fill*** brings you to a small dialogue screen where you can enter a formula (Fig. 6.5)

- 1…24 is called a **progression**; a list of numbers ascending with equal increments or descending with equal decrements. 1…24 returns 1, 2, 3, 4, 5 up to 24. A second number separated from the first by a comma gives the increment or decrement. 1,2…24 would return 1, 3, 5, 7 and so on (24 would not be included).
- **Pre-multipliers** cause each number in a progression or list between brackets to be repeated. 6(1…4) would return 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4
- **Post-multipliers** cause the list to be repeated. (1…4)6 would return 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4

Fig. 6.5 The List Fill dialogue screen



Spreadshe...

Row	Variety	Yield
1	1	*
2	1	*
3	1	
4	1	
5	1	
6	1	
7	2	
8	2	
9	2	
10	2	
11	2	*

Fill Column Variety with a List

Column: Variety

List formulae: 6(1...4)

Fill from cell: ● Top ○ Current

Fill to: ● Bottom ○ End of list

Current cell to fill from: 1

Again, convert the column to a factor afterwards and change the labels.

➔ **Converting text columns.**

This option is very useful when importing for instance survey data from other software packages. On a text column you can right-click and choose ***Convert to Factor***. The result is a factor column in GenStat containing labels.

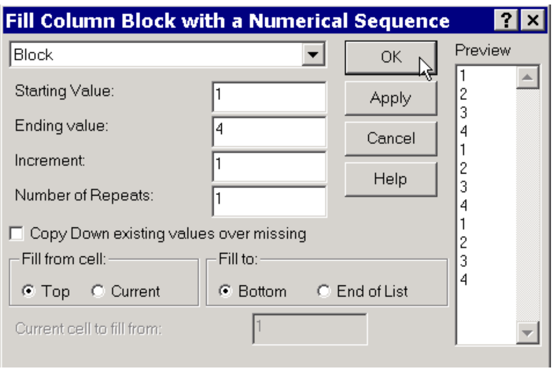
6.1.1.2 Layers and light regimes.

Use any of the methods mentioned in chapter 2.3.3 and chapter 6.1.1.1 to create a spreadsheet with data of an experiment on the effect of lighting on the egg production of pullets (Mead, Curnow and Hasted page 69). The figures are the number of eggs laid by a pen of six pullets in the period between 1 December 1950 and 22 February 1951.

Blocks	1	2	3	4	Treatment
Treatments					totals
O	330	288	295	313	1226
E	372	340	343	341	1396
F	359	337	373	302	1371
Block totals	1061	965	1011	956	3993

If you use the ***List fill*** option, note that the formula to enter the blocks will be (1…4)3 while the formula for the treatments will be 4(1…3). The difference when using ***Spread => Calculate => Fill*** is shown in Fig. 6.6 and Fig. 6.7.

Fig. 6.6 The Fill dialogue screen when entering the blocks



Fill Column Block with a Numerical Sequence

Block

Starting Value: 1

Ending value: 4

Increment: 1

Number of Repeats: 1

☐ Copy Down existing values over missing

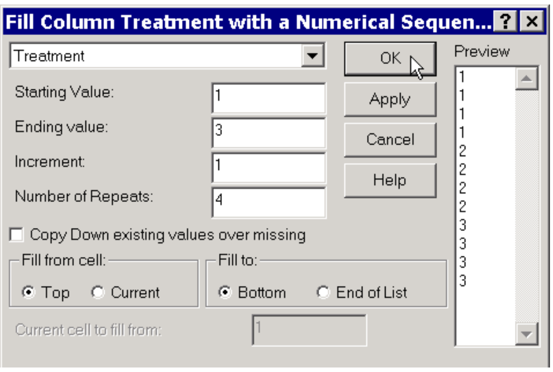
Fill from cell: ● Top ○ Current

Fill to: ● Bottom ○ End of List

Current cell to fill from: 1

Preview: 1 2 3 4 1 2 3 4 1 2 3 4

Fig. 6.7 The Fill dialogue screen when entering the treatments



Fill Column Treatment with a Numerical Sequen...

Treatment

Starting Value: 1

Ending value: 3

Increment: 1

Number of Repeats: 4

☐ Copy Down existing values over missing

Fill from cell: ● Top ○ Current

Fill to: ● Bottom ○ End of List

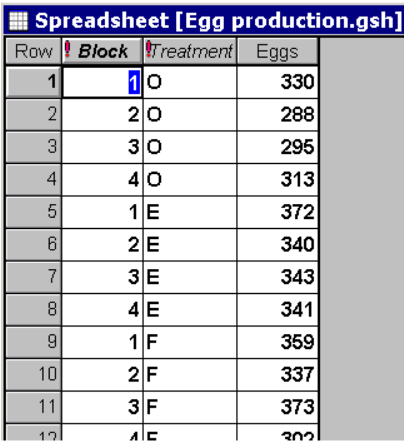
Current cell to fill from: 1

Preview: 1 1 1 1 2 2 2 2 3 3 3 3

The resulting GenStat spreadsheet should look as in Fig. 6.8. The following information can be included as column attributes to the treatment column.

- Treatment O: control (natural daylight only)
- Treatment E: extended day (total day length 14 hours)
- Treatment F: flash lightning (natural day plus twice 20 second flashes per night)

Fig. 6.8 The resulting spreadsheet



Spreadsheet [Egg production.gsh]

Row	Block	Treatment	Eggs
1	1	O	330
2	2	O	288
3	3	O	295
4	4	O	313
5	1	E	372
6	2	E	340
7	3	E	343
8	4	E	341
9	1	F	359
10	2	F	337
11	3	F	373
12	4	F	302

Save the file as “*Egg production.gsh*”.

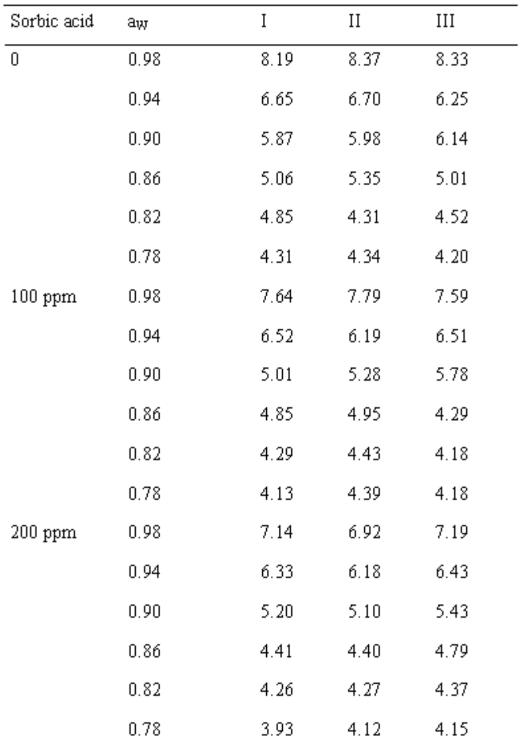
6.1.2 From an Excel spreadsheet.

6.1.2.1 Survival of Salmonella typhimurium.

The layout of the data where each factor or measurement is entered in a single column, as introduced in chapter 6.1.1, is the most suitable layout for working with any statistical package. However, if you have many factors or measurements, it can be difficult to understand the contents of a column. One way of documenting the data in GenStat is to give an extra description as was seen in chapter 2.2.1

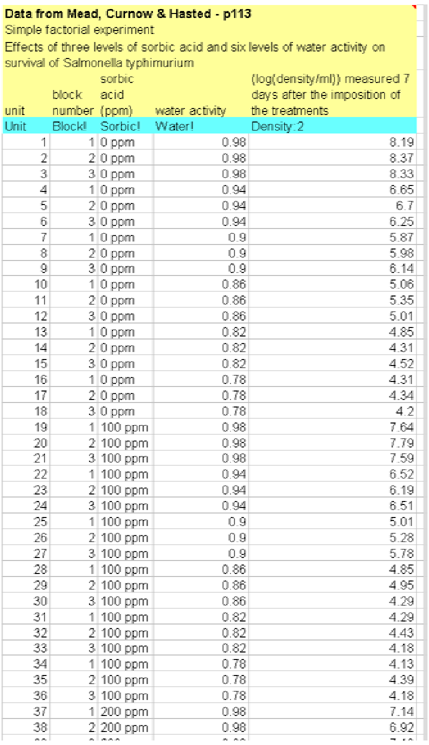
A more convenient way of adding a description to factors and measurements and also adding meta-documentation about the dataset can be done using Excel, as shown in the example below. Fig. 6.9 shows the textbook layout, while Fig. 6.10 shows the layout as entered in the Excel file “*Salmonella typhirum survival.xls*”. The data are from Mead, Curnow and Hasted, 2003 page 113.

Fig. 6.9 Text book layout of data



Sorbic acid	aw	I	II	III
0	0.98	8.19	8.37	8.33
	0.94	6.65	6.70	6.25
	0.90	5.87	5.98	6.14
	0.86	5.06	5.35	5.01
	0.82	4.85	4.31	4.52
	0.78	4.31	4.34	4.20
100 ppm	0.98	7.64	7.79	7.59
	0.94	6.52	6.19	6.51
	0.90	5.01	5.28	5.78
	0.86	4.85	4.95	4.29
	0.82	4.29	4.43	4.18
	0.78	4.13	4.39	4.18
200 ppm	0.98	7.14	6.92	7.19
	0.94	6.33	6.18	6.43
	0.90	5.20	5.10	5.43
	0.86	4.41	4.40	4.79
	0.82	4.26	4.27	4.37
	0.78	3.93	4.12	4.15

Fig. 6.10 Meta-documentation layout of data in MS Excel



Data from Mead, Curnow & Hasted - p113

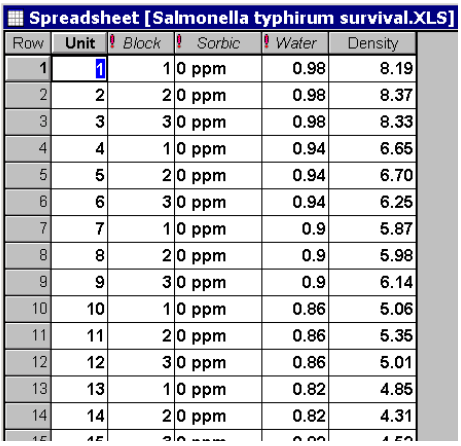
Simple factorial experiment

Effects of three levels of sorbic acid and six levels of water activity on survival of Salmonella typhimurium

unit	block number	sorbic acid (ppm)	water activity	(log(density/ml)) measured 7 days after the imposition of the treatments
Unit	Block!	Sorbic!	Water!	Density:2
1	1	0 ppm	0.98	8.19
2	2	0 ppm	0.98	8.37
3	3	0 ppm	0.98	8.33
4	1	0 ppm	0.94	6.65
5	2	0 ppm	0.94	6.7
6	3	0 ppm	0.94	6.25
7	1	0 ppm	0.9	5.87
8	2	0 ppm	0.9	5.98
9	3	0 ppm	0.9	6.14
10	1	0 ppm	0.86	5.06
11	2	0 ppm	0.86	5.35
12	3	0 ppm	0.86	5.01
13	1	0 ppm	0.82	4.85
14	2	0 ppm	0.82	4.31
15	3	0 ppm	0.82	4.52
16	1	0 ppm	0.78	4.31
17	2	0 ppm	0.78	4.34
18	3	0 ppm	0.78	4.2
19	1	100 ppm	0.98	7.64
20	2	100 ppm	0.98	7.79
21	3	100 ppm	0.98	7.59
22	1	100 ppm	0.94	6.52
23	2	100 ppm	0.94	6.19
24	3	100 ppm	0.94	6.51
25	1	100 ppm	0.9	5.01
26	2	100 ppm	0.9	5.28
27	3	100 ppm	0.9	5.78
28	1	100 ppm	0.86	4.85
29	2	100 ppm	0.86	4.95
30	3	100 ppm	0.86	4.29
31	1	100 ppm	0.82	4.29
32	2	100 ppm	0.82	4.43
33	3	100 ppm	0.82	4.18
34	1	100 ppm	0.78	4.13
35	2	100 ppm	0.78	4.39
36	3	100 ppm	0.78	4.18
37	1	200 ppm	0.98	7.14
38	2	200 ppm	0.98	6.92

As mentioned in chapter 2.2.2, you can define a name in Excel for the range containing only the data plus the header row, and import the named range into GenStat. Part of the GenStat spreadsheet after importing the Excel named range is shown in Fig. 6.11.

Fig. 6.11 Part of an imported Excel spreadsheet



Spreadsheet [Salmonella typhirum survival.XLS]

Row	Unit	! Block	! Sorbic	! Water	Density
1	1	1	0 ppm	0.98	8.19
2	2	2	0 ppm	0.98	8.37
3	3	3	0 ppm	0.98	8.33
4	4	1	0 ppm	0.94	6.65
5	5	2	0 ppm	0.94	6.70
6	6	3	0 ppm	0.94	6.25
7	7	1	0 ppm	0.9	5.87
8	8	2	0 ppm	0.9	5.98
9	9	3	0 ppm	0.9	6.14
10	10	1	0 ppm	0.86	5.06
11	11	2	0 ppm	0.86	5.35
12	12	3	0 ppm	0.86	5.01
13	13	1	0 ppm	0.82	4.85
14	14	2	0 ppm	0.82	4.31

Some tricks for easily importing a named range from Excel into GenStat.

An **exclamation mark (!)** following the column header in Excel (for example "Sorbic!") automatically converts the column to be a factor column in GenStat. In the resulting GenStat spreadsheet, the name of the column will be in italics and will be preceded by an exclamation mark.

A **dollar sign ($)** following the column header automatically converts the data in the column to text. In the resulting GenStat spreadsheet, a green letter T will precede the name of the column.

A **colon (:)** followed by a figure behind the column header formats the variate automatically to a certain number of decimals. For instance when importing the column Density:2 from Excel into GenStat, that variate will be formatted with two decimals.

Column headers (called **identifiers** in GenStat) can consist of up to 32 letters or digits, but they must start with a letter and they are case sensitive. Spaces between several words are converted to an underscore. However, avoid using long names or special characters (@, #, /, …) since this can create problems when exporting to other software packages.

A description of the column can be added in the row above the column name.

If any values are missing, an asterisk (*) could be entered. Blank Excel cells are automatically converted into an asterisk, when importing into GenStat. However both blank cells and asterisks can create confusion when exporting to other software packages. Make sure you distinguish blank cells from cells containing the value zero.

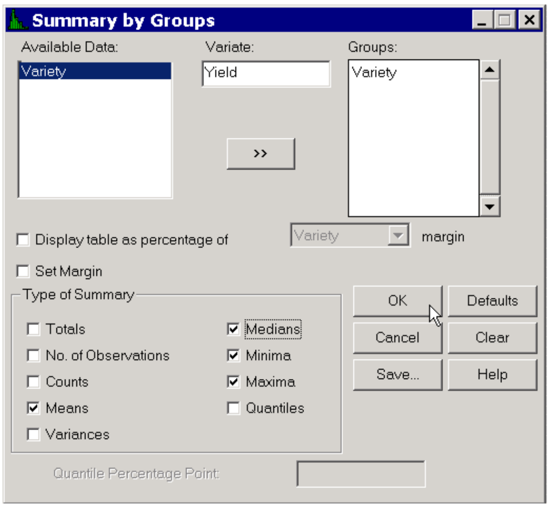
6.2 Exploratory analysis

> Before proceeding to the Analysis of Variance it is important to look (critically) at the data, both for error checking and to see if we can discover patterns in the data. We will do this now for the 3 spreadsheets we have created.

6.2.1 Melon yields

In chapter 2.3.1 we saw already how to create summary statistics. As always there are still other ways. Clear the GenStat memory (Run => Restart Session), open the file “*Melon yield.gsh*” and choose ***Stats => Summary Statistics => Summaries of Groups (Tabulation)***. Complete the dialogue as shown in Fig. 6.12.

Fig. 6.12 The tabulation dialogue window



Summary by Groups

Available Data: Variety

Variate: Yield

Groups: Variety

□ Display table as percentage of Variety margin

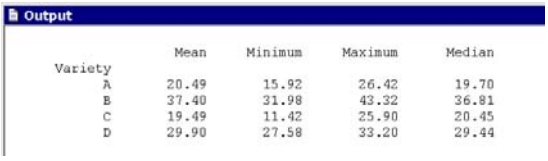
□ Set Margin

Type of Summary

□ Totals
□ No. of Observations
□ Counts
☑ Means
□ Variances
☑ Medians
☑ Minima
☑ Maxima
□ Quantiles

Quantile Percentage Point:

Fig. 6.13 The resulting table in the Output Window



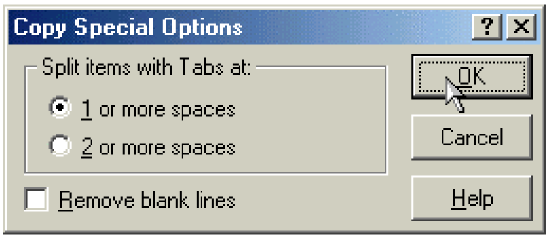
Output

```
                  Mean      Minimum      Maximum       Median
   Variety
         A        20.49        15.92        26.42        19.70
         B        37.40        31.98        43.32        36.81
         C        19.49        11.42        25.90        20.45
         D        29.90        27.58        33.20        29.44
```

The results appear in the output window (Fig. 6.13). You may now want to copy these summary statistics in a report. The following options will save you time.

Highlight the table in the Output Window and choose ***Edit => Copy Special => RTF Table***. Two dialogue boxes will appear. The first asks you after how many spaces you want to split the columns (Fig. 6.14); the second allows you to change the appearance of the final Word table (Fig. 6.15). Modify the options to suit the style of your report and click ***[OK]***.

Fig. 6.14 Splitting columns after one or two spaces



Copy Special Options

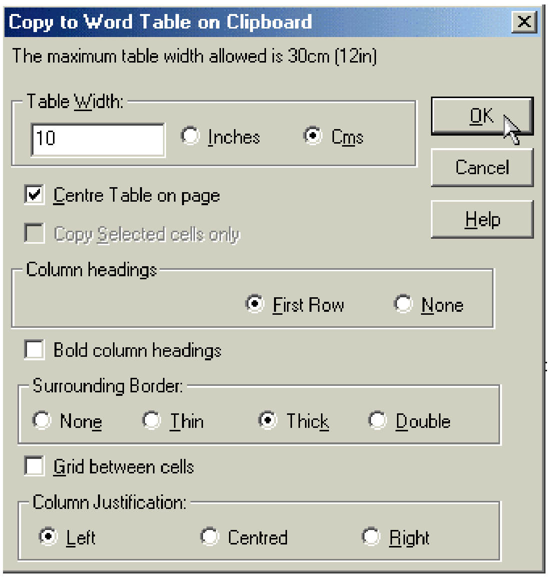
Split items with Tabs at:

● 1 or more spaces

○ 2 or more spaces

□ Remove blank lines

Fig. 6.15 Options for a RTF table



Copy to Word Table on Clipboard

The maximum table width allowed is 30cm (12in)

Table Width: 10 ○ Inches ● Cms

☑ Centre Table on page

□ Copy Selected cells only

Column headings: ● First Row ○ None

□ Bold column headings

Surrounding Border: ○ None ○ Thin ● Thick ○ Double

□ Grid between cells

Column Justification: ● Left ○ Centred ○ Right

The table below is what you get when you choose ***Edit => Paste Cells*** in Word. The resulting table looks as follows:

Mean	Minimum	Maximum	Median	
Variety				
A	20.49	15.92	26.42	19.70
B	37.40	31.98	43.32	36.81
C	19.49	11.42	25.90	20.45
D	29.90	27.58	33.20	29.44

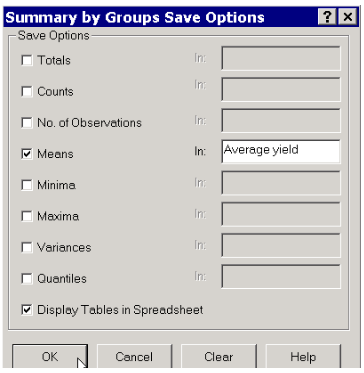
Notice that the headings of the columns are not correct. You could have edited the table in the Output Window of GenStat before copying it. Or you can modify the layout of the table in Word. The example below gives the same table with some columns deleted, some changes in the layout and some extra information added.

Melon variety	Average yield (kg)
A	20.49
B	37.40
C	19.49
D	29.90

Data from Mead, Curnow and Hasted, 2003. p. 58

Alternatively, you could first save your summary statistics in a GenStat spreadsheet. First click on the ***[Save]*** button in the Summary by Groups dialogue window. Check the summary statistics you want and give them a name. In the example we want the means and will call this “*Average yield*”. Don’t forget to indicate you want to show the resulting table in a spreadsheet (Fig. 6.16). After clicking ***[OK]*** a new table containing the summary statistics will appear (Fig. 6.17). You can format the means to have 2 decimals (see chapter 2.3.2).

Fig. 6.16 Saving the table

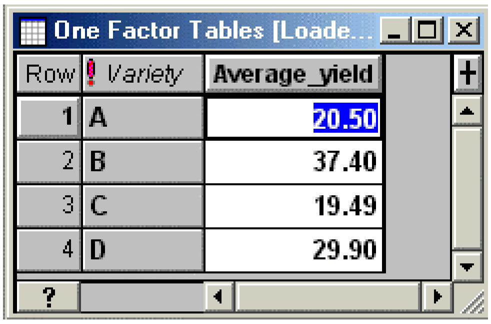


Summary by Groups Save Options

Save Options

- ☐ Totals In:
- ☐ Counts In:
- ☐ No. of Observations In:
- ☑ Means In: Average yield
- ☐ Minima In:
- ☐ Maxima In:
- ☐ Variances In:
- ☐ Quantiles In:
- ☑ Display Tables in Spreadsheet

Fig. 6.17 The resulting GenStat table



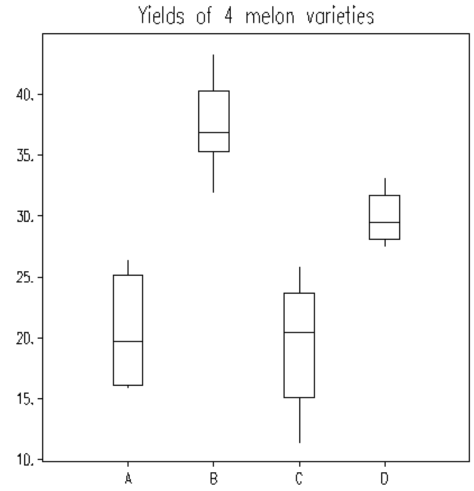
One Factor Tables [Loade...

Row	Variety	Average_yield
1	A	20.50
2	B	37.40
3	C	19.49
4	D	29.90

With the table of means selected, you can now choose again ***Edit => Copy Special => RTF Table.*** The next dialogue window is the same as in Fig. 6.15 and allows you to change the appearance of the final Word table. Again, you can paste the cells in Word and modify the layout of the table.

From the resulting table we clearly see that the average yield of variety B is higher than the others. Variety A and C have a low average yield, probably not much different. We cannot see if these differences are caused by a few observations or if the averages give a picture of the average situation. For this a boxplot is useful. Refer to chapter 3.2.1 (page 38) to create a similar boxplot to the one in Fig. 6.18.

Fig. 6.18 Boxplot of the yields of the melon varieties

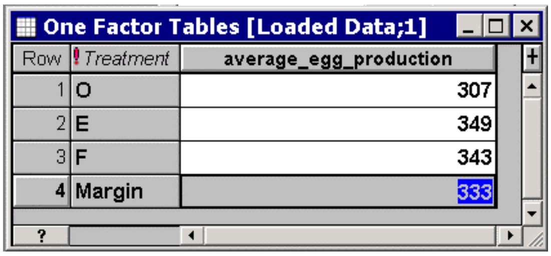


The exploratory analysis gives us an idea of which melon varieties have a higher yield and which ones produce a similar yield. Whether these are real differences will be shown later during the formal statistical analysis.

6.2.2 Layers and light regimes.

Follow the steps described in the previous section with the file "Egg production.gsh", to give the results shown in Fig. 6.19 and Fig. 6.20.

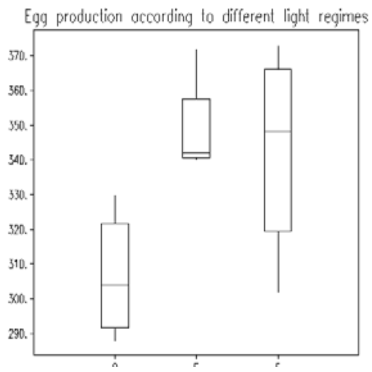
Fig. 6.19 GenStat table with average egg production per treatment



One Factor Tables [Loaded Data;1]

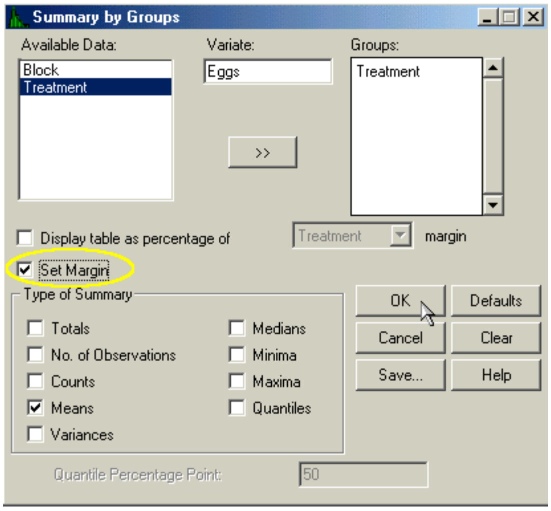
Row	Treatment	average_egg_production
1	O	307
2	E	349
3	F	343
4	Margin	333

Fig. 6.20 The resulting boxplot



The “Margin” in Fig. 6.19 which can be renamed as “Grand Mean”, is given by selecting the ***Set Margin*** box in the Summary by Groups dialogue window (Fig. 6.21) before you click on the ***[Save]*** button.

Fig. 6.21 Setting the margin



Summary by Groups

Available Data: Block, Treatment

Variate: Eggs

Groups: Treatment

☐ Display table as percentage of Treatment margin

☑ Set Margin

Type of Summary

☐ Totals ☐ Medians
☐ No. of Observations ☐ Minima
☐ Counts ☐ Maxima
☑ Means ☐ Quantiles
☐ Variances

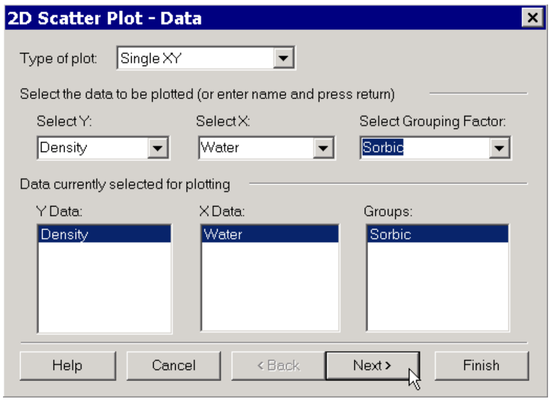
Quantile Percentage Point: 50

6.2.3 Survival of Salmonella typhimurium.

Open “*Salmonella typhirum survival.xls*” and perform some exploratory analysis. In some cases, a scatter plot (also called a point plot), is more useful than boxplots to draw conclusions. One question here is if there is a change in Salmonella density with increasing water acidity for the different levels of sorbic acid.

Use ***Graphics ⇒ Point plot*** and complete the dialogue as shown in Fig. 6.22.

Fig. 6.22 The scatter plot dialogue window



2D Scatter Plot - Data

Type of plot: Single XY

Select the data to be plotted (or enter name and press return)

Select Y: Density

Select X: Water

Select Grouping Factor: Sorbic

Data currently selected for plotting

Y Data: Density

X Data: Water

Groups: Sorbic

Fig. 6.23 The resulting graph

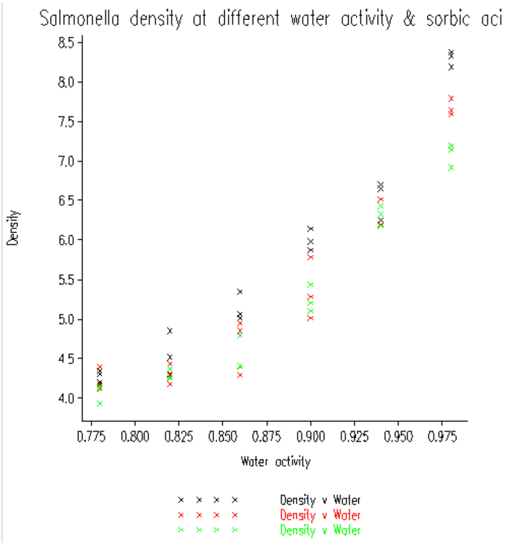
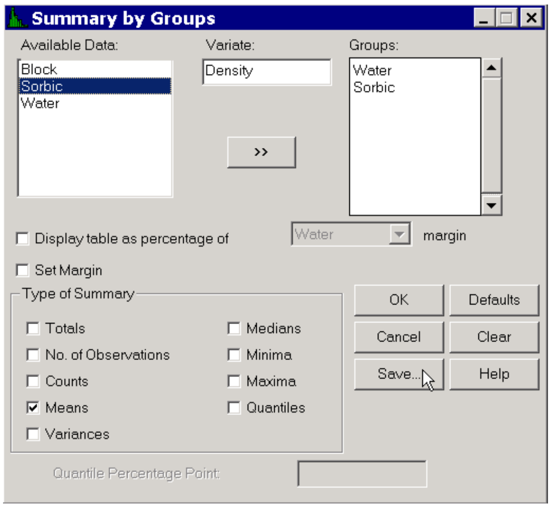


Fig. 6.23 shows a clear pattern of increasing Salmonella density with increasing water activity. In the GenStat Discovery Edition, it is however not easy to see the different levels of sorbic acid. We see they have different colours, but we don’t learn more from the legend. In the latest version, it is possible to change legends, titles and axes after a graph has been plotted.

A possible way to work around this problem with the Discovery Edition is to plot the group averages. Choose ***Stats => Summary Statistics => Summaries of Groups (Tabulation)*** and calculate the means of Density by Water activity and Sorbic acid level (Fig. 6.24). Click on the ***[Save]*** button to save the means in a table (Fig. 6.25). This table will show the levels of the factor that is put at the bottom in the Groups box (Fig. 6.24) as columns. The other factors will be shown as rows.

Fig. 6.24 First step while creating a table of average Density by Water activity and Sorbic acid level



Summary by Groups

Available Data: Block, Sorbic, Water

Variate: Density

Groups: Water, Sorbic

□ Display table as percentage of Water margin

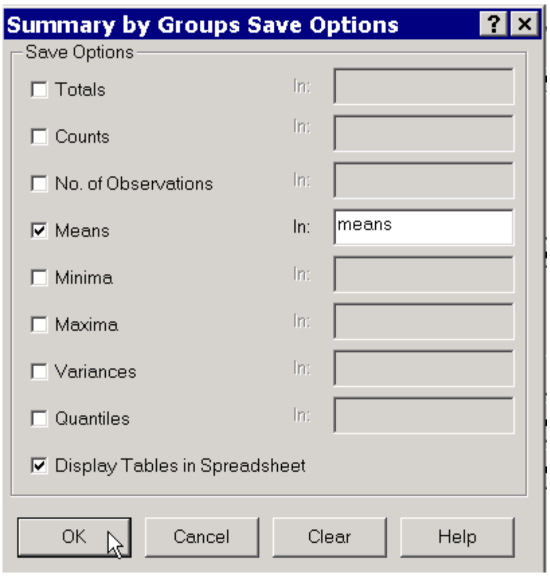
□ Set Margin

Type of Summary

□ Totals
□ No. of Observations
□ Counts
☑ Means
□ Variances
□ Medians
□ Minima
□ Maxima
□ Quantiles

Quantile Percentage Point

Fig. 6.25 Finishing the table



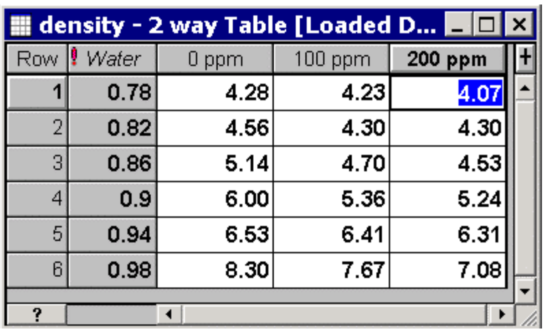
Summary by Groups Save Options

Save Options

□ Totals In:
□ Counts In:
□ No. of Observations In:
☑ Means In: means
□ Minima In:
□ Maxima In:
□ Variances In:
□ Quantiles In:
☑ Display Tables in Spreadsheet

The resulting spreadsheet is a table shown in Fig. 6.26. This is another GenStat data structure. It is not possible to plot a graph of the data in a table, but we can convert the table to the normal type of spreadsheet (the “Vector” type). Click in the table shown in Fig. 6.26 and choose ***Spread => Manipulate => Convert***, indicate you want the sheet type “***Vector***” and click ***[OK]*** (see Fig. 6.27). Also, format the columns to show 2 decimals.

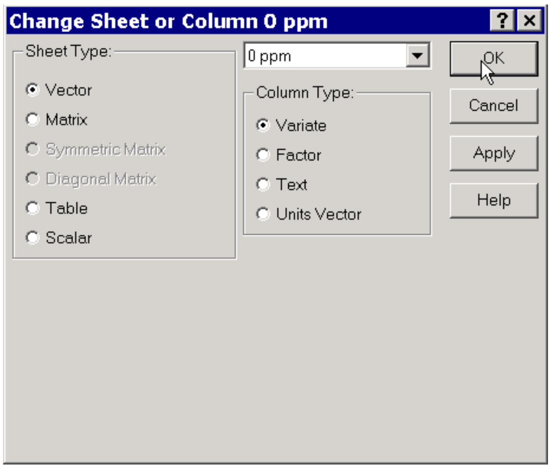
Fig. 6.26 A GenStat table



density - 2 way Table [Loaded D...

Row	Water	0 ppm	100 ppm	200 ppm
1	0.78	4.28	4.23	4.07
2	0.82	4.56	4.30	4.30
3	0.86	5.14	4.70	4.53
4	0.9	6.00	5.36	5.24
5	0.94	6.53	6.41	6.31
6	0.98	8.30	7.67	7.08

Fig. 6.27 Converting the table to a vector sheet type



Change Sheet or Column 0 ppm

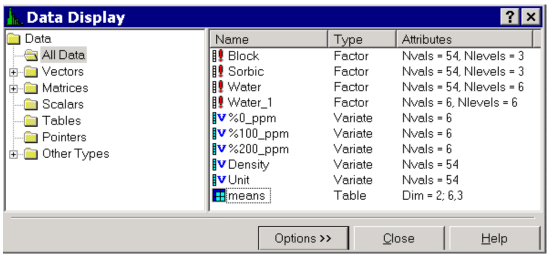
Sheet Type:
● Vector
○ Matrix
○ Symmetric Matrix
○ Diagonal Matrix
○ Table
○ Scalar

0 ppm

Column Type:
● Variate
○ Factor
○ Text
○ Units Vector

If you now check the available data in the GenStat server (see Fig. 6.28 as was shown in chapter 2.4.1), you see you have created three new variates and one new factor, each of them with 6 values. GenStat converted the column headers from the table into those variates and factors but at the same time changed their name (Water_1, %0_ppm, %100_ppm, %200_ppm). We will not go into details about this naming. You can now create a point plot. Since the y-values of such a plot are in 3 different variates, we choose a Multiple Y Scatter Plot in Fig. 6.29.

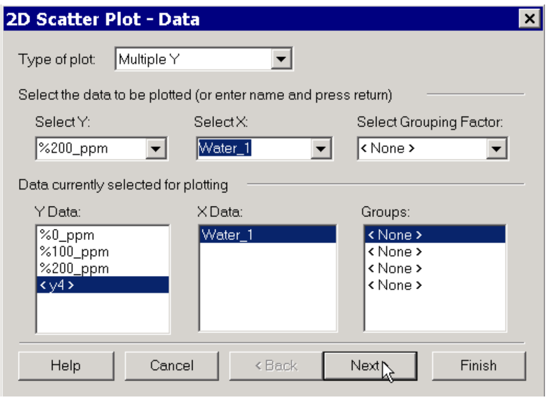
Fig. 6.28 New data structures in the available data



Data Display

Name	Type	Attributes
Block	Factor	Nvals = 54, Nlevels = 3
Sorbic	Factor	Nvals = 54, Nlevels = 3
Water	Factor	Nvals = 54, Nlevels = 6
Water_1	Factor	Nvals = 6, Nlevels = 6
%0_ppm	Variate	Nvals = 6
%100_ppm	Variate	Nvals = 6
%200_ppm	Variate	Nvals = 6
Density	Variate	Nvals = 54
Unit	Variate	Nvals = 54
means	Table	Dim = 2; 6,3

Fig. 6.29 Creating a Multiple Y Scatter Plot



2D Scatter Plot - Data

Type of plot: Multiple Y

Select the data to be plotted (or enter name and press return)

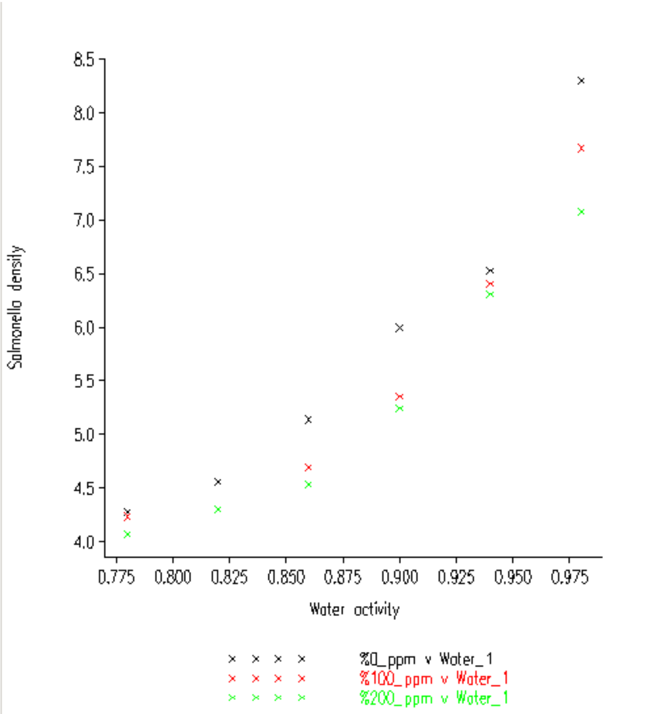
Select Y: %200_ppm
Select X: Water_1
Select Grouping Factor: < None >

Data currently selected for plotting

Y Data: %0_ppm, %100_ppm, %200_ppm, < y4 >
X Data: Water_1
Groups: < None >, < None >, < None >, < None >

The resulting scatter-plot of the group averages gives us a clear picture of the trend in the data set, and we can distinguish the different levels of Sorbic acid in the legend.

Fig. 6.30 The resulting scatter-plot with a clear legend



We conclude that the Salmonella density increases with increasing Water activity. The density also diminishes with increasing Sorbic acid. The question remains if these differences are significant. This is found out with a formal statistical analysis.

6.3 A real example.

Textbook examples are usually too small to demonstrate the importance of data exploration. A “real” example is therefore used as a further illustration of exploratory methods. This was an experiment in Kenya with 16 farmers in one district and 12 in a second. Each farmer had 3 plots, two of which had tree mulch applied; the third was a control. The main variable of interest was the grain yield. The primary objective of this experiment was to investigate whether the “good results” with the Tithonia and Lantana mulch in on-station experiments was repeated on farmer’s fields.

Import the named range “*data*” from the Excel file “*Onfarm tithonia and lantana mulches.xls*”. One topic of interest was whether all farmers from the West district benefited from the mulch in the experiment.

There is a spreadsheet menu ‘***Restrict***’ that is very useful when exploring data. Choose ***Spread => Restrict/Filter***. There are several possibilities and we choose to restrict by selecting a factor level (Fig. 6.31). In this case we want to select only the farmers of the West district and *West* is one of the two levels of the factor *location*. So we restrict the data to include only the level ‘*West*’ of the factor ‘*location*’ as in Fig. 6.32.

Fig. 6.31 Restricting to a factor level

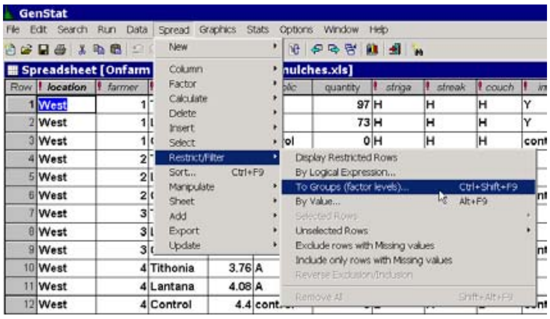
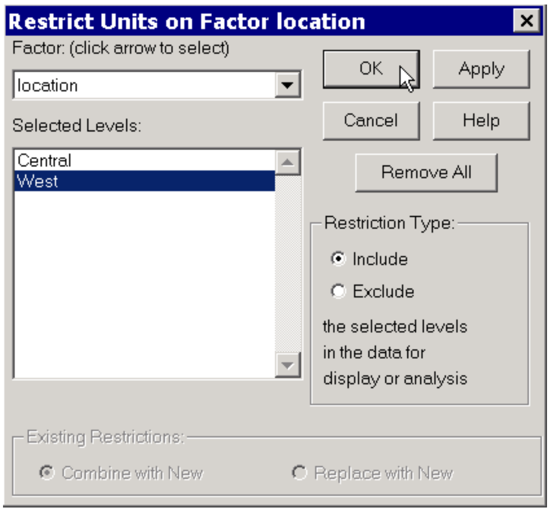

GenStat
File Edit Search Run Data Spread Graphics Stats Options Window Help
Spreadsheet [Onfarm ... mulches.xls]
Row location farmer quantity striga streak couch
New
Column
Factor
Calculate
Delete
Insert
Select
Restrict/Filter
Sort... Ctrl+F9
Manipulate
Sheet
Add
Export
Update
Display Restricted Rows
By Logical Expression...
To Groups (factor levels)... Ctrl+Shift+F9
By Value... Alt+F9
Selected Rows
Unselected Rows
Exclude rows with Missing values
Include only rows with Missing values
Reverse Exclusion/Inclusion
Remove All Shift+Alt+F9

Fig. 6.32 Selecting only the location ‘West’


Restrict Units on Factor location

Factor: (click arrow to select) location

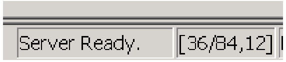
Selected Levels:
Central
West

Restriction Type:
● Include
○ Exclude
the selected levels in the data for display or analysis

Existing Restrictions:
● Combine with New ○ Replace with New

The result of this restriction is that all data from the Central district are still there but they are not used in the calculations. The status bar (Fig. 6.33) shows for instance that only 36 of the 84 rows are not restricted.

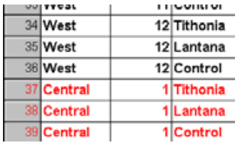
Fig. 6.33 The status bar shows the number of non-restricted rows.


Server Ready. [36/84,12]

When you click on the ‘restrict switch’ (Fig. 6.34); a cross above the scroll bar of the spreadsheet, the restricted rows will be displayed in red as shown in Fig. 6.35.

Fig. 6.34 The restrict switch

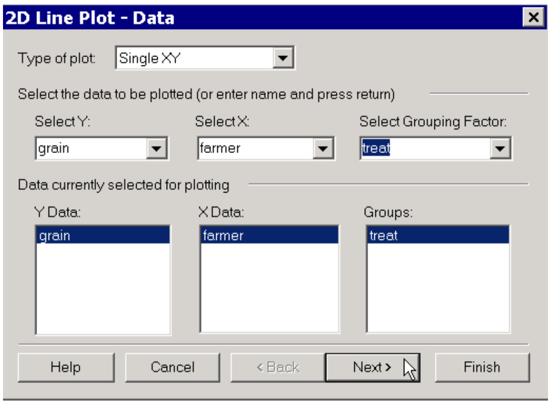
Fig. 6.35 Restricted rows are shown in red



34	West	12	Tithonia
35	West	12	Lantana
36	West	12	Control
37	Central	1	Tithonia
38	Central	1	Lantana
39	Central	1	Control

To see if farmers in the West district who applied mulch obtained a higher maize yield, we could construct a line plot. Choose ***Graphics => Line plot*** and construct a plot of maize yield (the variate *grain*) versus *farmer* grouped per type of mulch (variate *treat*).

Fig. 6.36 Constructing a line plot



2D Line Plot - Data

Type of plot: Single XY

Select the data to be plotted (or enter name and press return)

Select Y: grain

Select X: farmer

Select Grouping Factor: treat

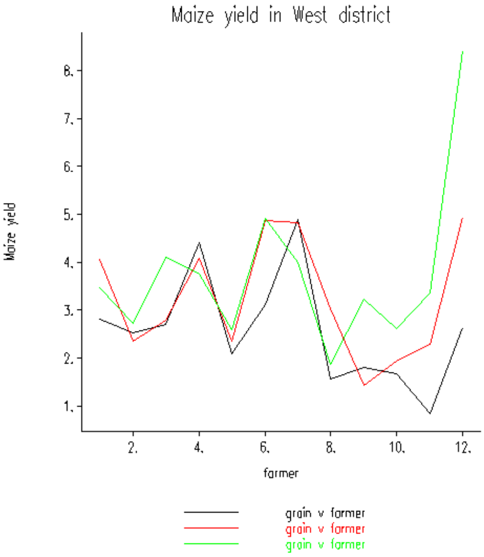
Data currently selected for plotting

Y Data: grain

X Data: farmer

Groups: treat

Fig. 6.37 The resulting line plot

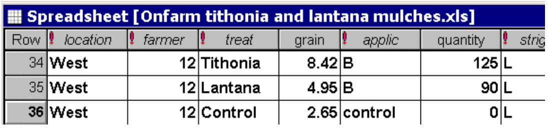


The resulting graph (Fig. 6.37) helps but not much. First of all, because we are working in an older version of the GenStat graphics, the legend shows '*grain versus farmer*' three times but in a different colour. We cannot see which type of mulch gives the highest yield. The second problem is that the data along the X-axis are organised with increasing farmer number. It would be more helpful if the data were organised with increasing average grain yield.

We can work around the first problem by using the latest version of GenStat, or by using some tricks in GenStat Discovery Edition and MS Word. From the graph in Fig. 6.37 it is obvious that the highest yield for farmer 12 is given by the treatment symbolized with a green line. Then follows the red line and the lowest yield is given by the black line.

To find which mulches these are you could for instance further restrict the restricted dataset, but this time '***To Groups(factor levels)***' farmer is 12 and make sure to select that you combine this restriction with the existing restrictions.

Fig. 6.38 The resulting spreadsheet after a second restriction



Spreadsheet [Onfarm tithonia and lantana mulches.xls]

Row	location	farmer	treat	grain	applic	quantity	strig
34	West	12	Tithonia	8.42	B	125	L
35	West	12	Lantana	4.95	B	90	L
36	West	12	Control	2.65	control	0	L

In the resulting spreadsheet of Fig. 6.38 you see that the Tithonia mulch gave the highest maize yield and the control the lowest. So, the green line is Tithonia mulch, the red line is Lantana mulch and the black line is no mulch (control).

To use this graph in a report, save it as a bitmap file. In the GenStat 4.1 graphics window, choose ***File => Save as***, set the file type to bitmap file and name the file for instance '*mulch.bmp*'. Now open MS Word. In MS Word, choose ***Insert => Picture => From File*** and insert the mulch.bmp file. Now select the drawing toolbar (***View => Toolbars => Drawing***) and click on the text box button (Fig. 6.39). Type "Control" in the first text box and drag it over the text next to the black line in the legend as in Fig. 6.40.

Fig. 6.39 Creating a Text Box with the drawing toolbar in MS Word

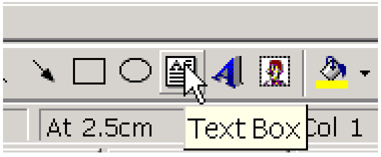
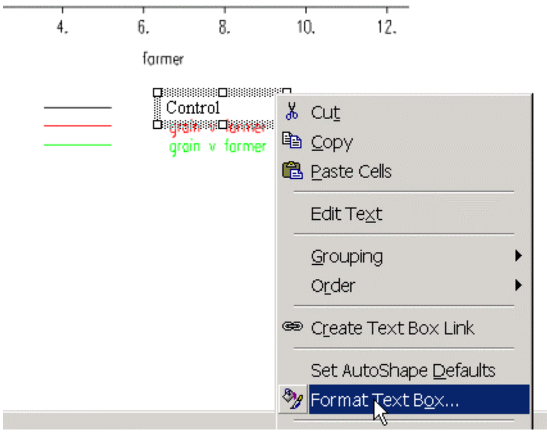


Fig. 6.40 Overwriting the GenStat legend with a text box in Word

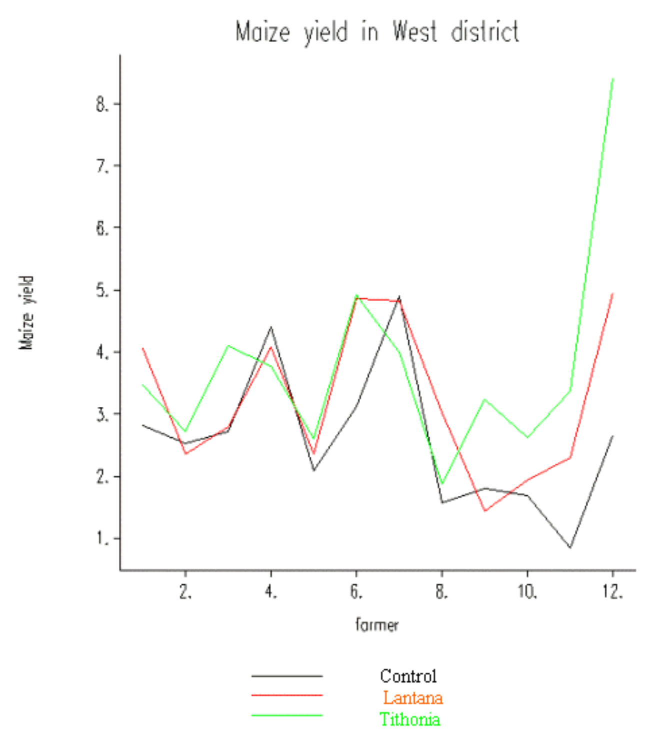


In Word, it will be necessary to format the text box. Right-click on it and choose ***Format Text Box***. You will probably have to:

- Change ***Colors and Lines***: choose a white fill colour and no border line.
- Make sure that the ***Wrapping style*** under the Layout tab is set "***In front of text***".
- The ***Internal margins*** under the ***Text box*** tab are set at 0 or 0.05 cm.
- Make sure that the font size of the text in the text box is not too big. Use for instance Times New Roman 9pt.

Repeat this for "Tithonia" and "Lantana", where you will also have to change the colour of the text. The resulting graph can be seen in Fig. 6.41.

Fig. 6.41 The resulting line plot with a clear legend



You could have done most of what we did using MS Word directly in GenStat by using the commands. If you look in the Input Window after having created the graph of Fig. 6.37, you see that GenStat used a whole bunch of commands with different parameters and options to create the graph as is shown below.

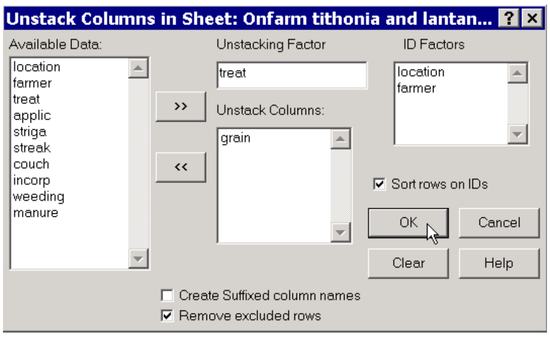
```
XAXIS [RESET=yes] WINDOW=1; TITLE='farmer'; TPOSITION=middle;\
TDIRECTION=parallel;LPOSITION=outside; LDIRECTION=parallel;\
MPOSITION=outside; ARROWHEAD=omit; ACTION=display
YAXIS [RESET=yes] WINDOW=1; TITLE='grain';\
TPOSITION=middle;TDIRECTION=parallel; LPOSITION=outside;\
LDIRECTION=perpendicular; MPOSITION=outside; ARROWHEAD=omit;\
ACTION=display
CALC _nlevs=NLEVELS( treat)
PEN [RESET=yes] 1..._nlevs; METHOD=line; JOIN=ascending;\
SYMBOL=0; LINESTYLE=1
DGRAPH [WINDOW=1; TITLE='Maize yield in West District'] Y=grain;\
X=farmer; PEN=NEWLEVELS( treat;!(1..._nlevs))
PEN [RESET=yes] 1..._nlevs
```

An alternative way to create a graph with a clear and unambiguous legend is to change some of the parameters and options and run the commands again. It is however not completely straightforward. Check the GenStat help for the commands XAXIS, YAXIS, PEN and DGRAPH.

In the graph in Fig. 6.41, the data along the X-axis are ordered according to farmer number. It would me more informative to order them according to increasing average maize yield. You could consider the column *grain* as a stack of maize yield data when there is no mulch, when Lantana mulch is applied and when Tithonia mulch is applied. With the Unstack menu command we can split the stack of data in three smaller stacks; in this case into one stack per type of mulch applied.

Choose ***Spread => Manipulate => Unstack*** (Fig. 6.42). We split the stack of maize (put *grain* in the ***Unstack Columns*** box) according to treatment (*treat* in the ***Unstacking Factor*** box) and we want to keep some other factors in the new spreadsheet, like *location* and *farmer*, to be able to identify the data values (put them in the ***ID Factors*** box). The result should look as in Fig. 6.43.

Fig. 6.42 The Unstack Columns dialogue box



Unstack Columns in Sheet: Onfarm tithonia and lantan...

Available Data: location, farmer, treat, applic, striga, streak, couch, incorp, weeding, manure

Unstacking Factor: treat

ID Factors: location, farmer

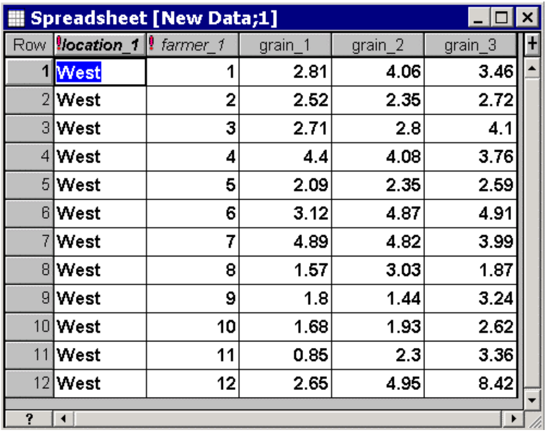
Unstack Columns: grain

☑ Sort rows on IDs

☐ Create Suffixed column names

☑ Remove excluded rows

Fig. 6.43 The resulting spreadsheet with unstacked columns

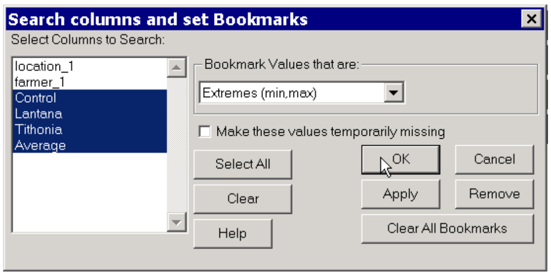


Spreadsheet [New Data;1]

Row	location_1	farmer_1	grain_1	grain_2	grain_3
1	West	1	2.81	4.06	3.46
2	West	2	2.52	2.35	2.72
3	West	3	2.71	2.8	4.1
4	West	4	4.4	4.08	3.76
5	West	5	2.09	2.35	2.59
6	West	6	3.12	4.87	4.91
7	West	7	4.89	4.82	3.99
8	West	8	1.57	3.03	1.87
9	West	9	1.8	1.44	3.24
10	West	10	1.68	1.93	2.62
11	West	11	0.85	2.3	3.36
12	West	12	2.65	4.95	8.42

Rename the columns grain_1, grain_2 and grain_3 according to the treatment (see chapter 2.2.1.1) and calculate the average grain yield in a new column (see chapter 2.3.2). With ***Search => Bookmark => By value***, you can now mark the minimum and maximum value of each variate (Fig. 6.44). When you right-click in the column with the average grain yield, a menu appears that allows you to sort the averages in ascending order (Fig. 6.45).

Fig. 6.44 Marking the extremes



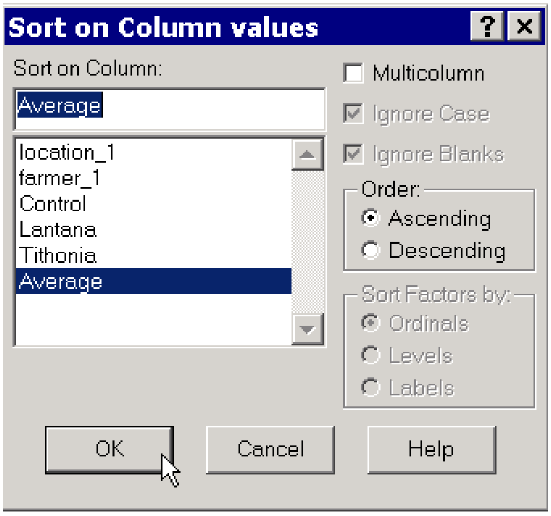
Search columns and set Bookmarks

Select Columns to Search: location_1, farmer_1, Control, Lantana, Tithonia, Average (selected: Control, Lantana, Tithonia, Average)

Bookmark Values that are: Extremes (min,max)

☐ Make these values temporarily missing

Fig. 6.45 Sorting on the average



Sort on Column values

Sort on Column: Average

location_1, farmer_1, Control, Lantana, Tithonia, Average

☐ Multicolumn

☑ Ignore Case

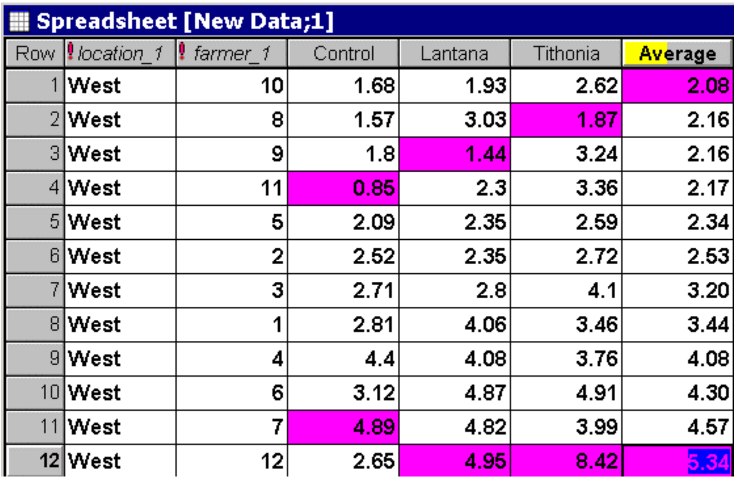
☑ Ignore Blanks

Order: ● Ascending ○ Descending

Sort Factors by: ● Ordinals ○ Levels ○ Labels

If you followed all steps above, you should get a spreadsheet that looks as in (Fig. 6.46)

Fig. 6.46 The sorted spreadsheet

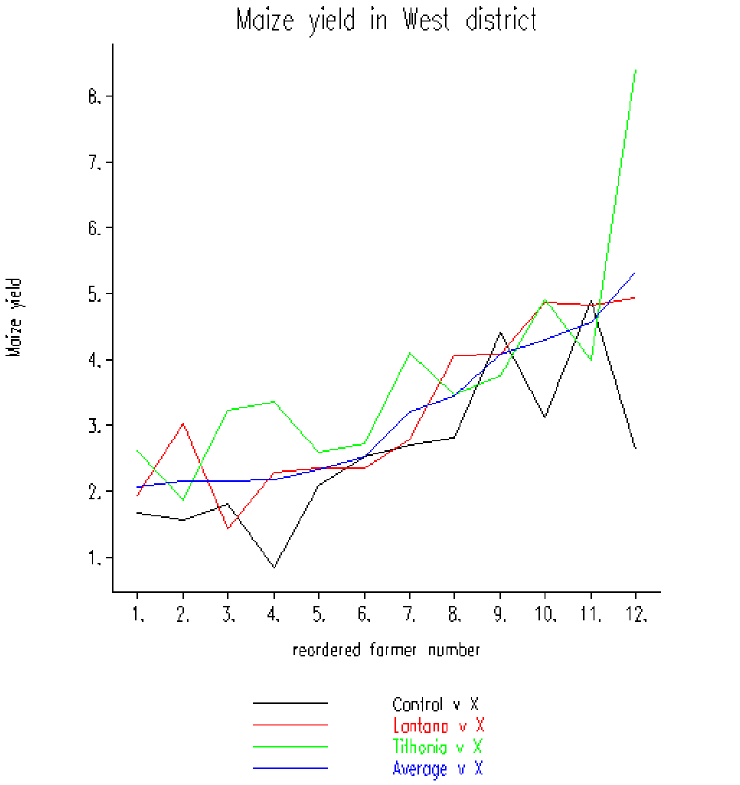


Spreadsheet [New Data;1]

Row	location_1	farmer_1	Control	Lantana	Tithonia	Average
1	West	10	1.68	1.93	2.62	2.08
2	West	8	1.57	3.03	1.87	2.16
3	West	9	1.8	1.44	3.24	2.16
4	West	11	0.85	2.3	3.36	2.17
5	West	5	2.09	2.35	2.59	2.34
6	West	2	2.52	2.35	2.72	2.53
7	West	3	2.71	2.8	4.1	3.20
8	West	1	2.81	4.06	3.46	3.44
9	West	4	4.4	4.08	3.76	4.08
10	West	6	3.12	4.87	4.91	4.30
11	West	7	4.89	4.82	3.99	4.57
12	West	12	2.65	4.95	8.42	5.34

Follow the steps mentioned above (but think about what we did in chapter 6.2.3 when choosing the type of plot). You should get a similar graph as the one in Fig. 6.47.

Fig. 6.47 The final graph with maize yield in ordered along the ascending average maize yield



Now back to the data exploration. Try to answer following questions, using either the table or the graph shown above:

- How many of the 12 farmers had a higher yield from the Tithonia mulch compared to the control ?

- And the Lantana mulch compared to the control?

In the original spreadsheet, use ***Spread*** ⇒ ***Restrict/Filter*** ⇒ ***Using Factor levels*** again and select only the ***Central*** district. Click to ***[Replace with New]***, otherwise you will have no data left! Use ***Graphics*** ⇒ ***Line Plot*** again to plot just the Central district data. This gives the equivalent plot for the Central district. Try to answer the same questions as above but for the Central district. From the two graphs do you feel that the farmers in the two districts benefit equally from the mulch?

> Remember that when you wish to use all the data, you must first use ***Spread*** ⇒ ***Restrict/Filter*** ⇒ ***Remove All.*** All calculations under the ***Stats*** menu will only be performed on the restricted data set. ***Bookmarks*** however will be set on all data.

7 Challenge 2

In the graph in Fig. 6.47 on page 78, the data along the X-axis are ordered according to increasing average maize yield. You became familiar with stacking and unstacking data in GenStat. There is also a second way of arriving at the graph in Fig. 6.47. Try to find it.

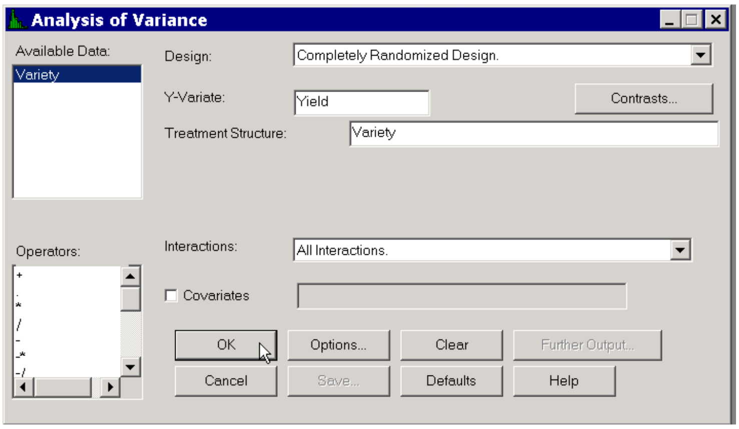
(Hint, use a tabulation first and set the margin.)

8 Analysis of variance.

8.1 Two simple ANOVA’s

Start a new GenStat session and open the file “*Melon yield.gsh*” (chapter 6.1.1.1 , page 60). The design of this experiment is a completely randomised design. From the exploratory analysis in chapter 6.2.1 we suspect there will be differences between the yields of different melon varieties. We confirm this with a formal statistical analysis. To perform an analysis of variance, choose ***Stats ⇒ Analysis of Variance.*** In the Analysis of Variance menu, choose ***Completely Randomized Design*** or ***One-Way ANOVA (no Blocking),*** see Fig. 8.1. Fill in the variate and the treatment and click ***[OK]***.

Fig. 8.1 The ANOVA dialogue window



Design: Completely Randomized Design.

Y-Variate: Yield

Treatment Structure: Variety

Interactions: All Interactions.

The resulting ANOVA table can be found in the Output Window.

```
***** Analysis of variance *****

Variate: Yield

Source of variation      d.f.       s.s.       m.s.    v.r.  F pr.
Variety                     3    1291.48     430.49   23.42  <.001
Residual                   20     367.65      18.38
Total                      23    1659.13
```

We conclude that there are significant differences in yield for the different melon varieties.

If you prefer to use the command language (see chapter 2.4.2), look in the Input Window to see which commands have been used.

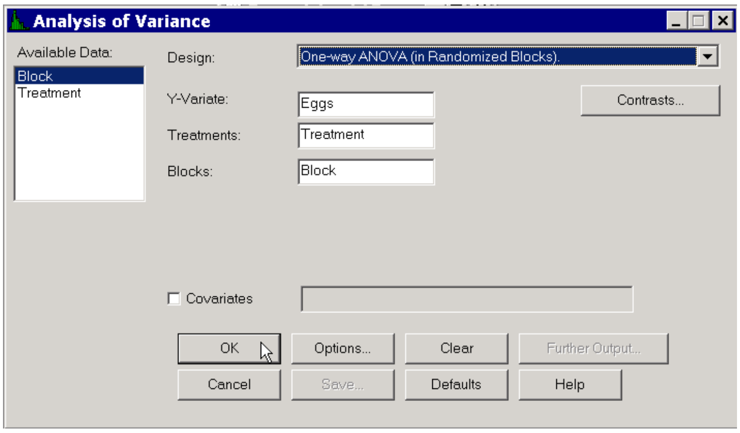
```
"Completely Randomized Design."
BLOCK "No Blocking"
TREATMENTS Variety
COVARIATE "No Covariate"
ANOVA [PRINT=aovtable,information,means; FACT=32; FPROB=yes;
PSE=diff] Yield
```

Restart the session and open “Egg production.gsh” (chapter 6.1.1.2 , page 62). This time there is not only a factor describing the treatments, but a second factor – *Block* – describing the layout (design) of the experiment.

To analyse data from such an experiment choose ***Stats ⇒ Analysis of Variance***, then select ***One-way ANOVA (in Randomised Blocks)*** from the list in the ***Design*** box, see Fig. 8.2.

The resulting dialogue box differs from a simple one-way ANOVA used earlier in that there is an extra box ***[Blocks]*** for the blocking factor. This allows information about the layout of the experiment to be passed to GenStat.

Fig. 8.2 The dialogue window for a one-way ANOVA in randomized blocks



Analysis of Variance

Available Data: Block, Treatment

Design: One-way ANOVA (in Randomized Blocks)

Y-Variate: Eggs

Treatments: Treatment

Blocks: Block

☐ Covariates

The results are shown in the Output Window. We conclude there are differences in egg production for the different treatments.

```
***** Analysis of variance *****

Variate: Eggs

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              3     2330.3      776.8    2.01

Block.*Units* stratum
Treatment                  2     4212.5     2106.3    5.44  0.045
Residual                   6     2321.5      386.9

Total                     11     8864.2
```

8.2 Getting more out of the output

The conclusion that there are significant differences between the treatments is just the starting point in the analysis. We now use the other information in the Output Window to assess what differences they are.

We expect that the treatments will increase the egg production. That is why we did the experiment. This increase can be calculated from the tables of means that can be found in the Output Window. Treatment F (flash lights) increases the number of eggs per 6 chickens by 36.3 (342.8 – 306.5) in the almost three month period (or by 36.3/6 = 6.05 eggs per chicken). Extended daylight increases the production by 42.5 eggs (i.e. 349 – 306.5); or 7.1 eggs per chicken.

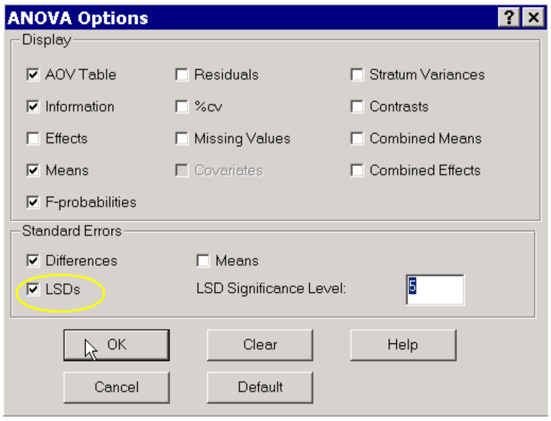
```
***** Tables of means *****

Variate: Eggs
 Grand mean  332.8
 Treatment        O        E        F
               306.5    349.0    342.8
```

The standard error of the differences of the means is also provided in the Output Window.

```
*** Standard errors of differences of means ***
Table            Treatment
rep.                     4
d.f.                     6
s.e.d.               13.91
```

The standard error times a t-value based on 6 degrees of freedom (the residual degrees of freedom from the ANOVA – also shown with the s.e.d.) is called the LSD (least significant difference). This can also be found in the Output Window by requesting it in the ANOVA Options dialogue window (Fig. 8.3). You get this window by clicking on the ***[Options…]*** button in the Analysis of Variance dialogue window.

Fig. 8.3 Changing the options what to include in the output of an ANOVA



Display

☑ AOV Table □ Residuals □ Stratum Variances
☑ Information □ %cv □ Contrasts
□ Effects □ Missing Values □ Combined Means
☑ Means □ Covariates □ Combined Effects
☑ F-probabilities

Standard Errors

☑ Differences □ Means
☑ LSDs LSD Significance Level: 5

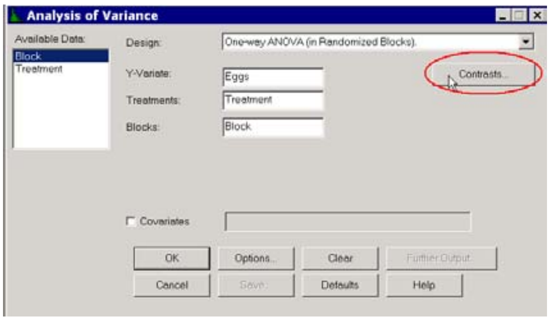
```
*** Least significant differences of means (5% level) ***

Table              Treatment
rep.                       4
d.f.                       6
l.s.d.                 34.03
```

So, we are fairly sure that either of the two treatments will increase the egg production compared to the control. Now, which of the two treatments is the “better”? The difference of 6.2 eggs per pen (or about 1 egg per chicken) is way below 34.03. Any difference between E and F is too small for this experiment to detect.

We have been using the standard output to compare the treatments. In this case, there were only a few treatments and the comparisons were done to answer the research questions: “Does extra light improve the egg production and which method is the better?”. We answered these questions and added a measure of precision.

One common way to ensure that the analysis corresponds to well-defined objectives is to use contrasts. We illustrate their use with this same example.

Fig. 8.4 The **[Contrasts]** *button in the ANOVA dialogue window*



Analysis of Variance

Available Data: Block, Treatment

Design: One-way ANOVA (in Randomized Blocks).

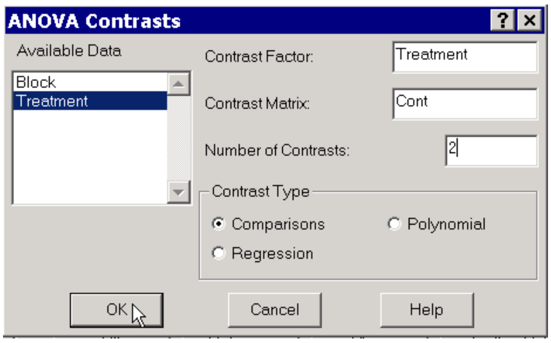
Y-Variate: Eggs

Treatments: Treatment

Blocks: Block

□ Covariates

Fig. 8.5 Specifying the contrast



ANOVA Contrasts

Available Data: Block, Treatment

Contrast Factor: Treatment

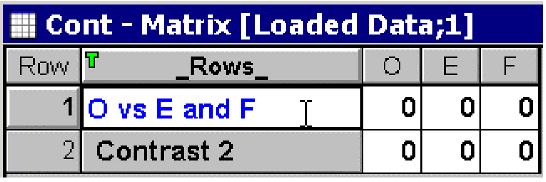
Contrast Matrix: Cont

Number of Contrasts: 2

Contrast Type: ● Comparisons ○ Polynomial ○ Regression

Click on the ***[Contrasts]*** button in the Analysis of Variance dialogue window (Fig. 8.4). A sub-dialogue window opens, where we specify the contrasts. We want to compare some treatments, so the *Contrast Factor* in the example is *Treatment* and the type of contrast is *Comparisons* (Fig. 8.5). We want to make 2 comparisons (light treatments versus control and extended daylight versus flashing), so the number of contrasts is 2. When you click [OK] you will get a matrix as shown in Fig. 8.6. The default name of the matrix is *Cont*, but you could have changed this in the ANOVA Contrasts dialogue window. This matrix has two rows, because we specified that we wanted to make 2 comparisons. It has 3 columns, because the treatment factor has 3 levels.

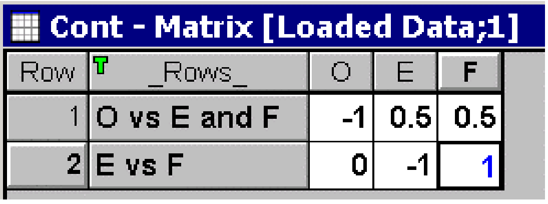
Fig. 8.6 Renaming the contrast



Cont - Matrix [Loaded Data;1]

Row	_Rows_	O	E	F
1	O vs E and F	0	0	0
2	Contrast 2	0	0	0

Fig. 8.7 Filling in the linear combination



Cont - Matrix [Loaded Data;1]

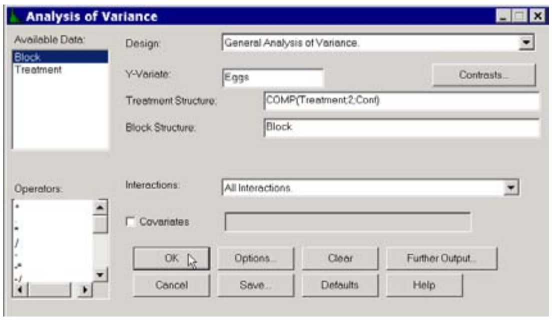
Row	_Rows_	O	E	F
1	O vs E and F	-1	0.5	0.5
2	E vs F	0	-1	1

By default the rows are labelled as “Contrast 1” and “Contrast 2” but you can change this by simply clicking in the cell. We are interested in 2 comparisons; “*O vs E and F*” and “*E vs F*”, see Fig. 8.7.

Now we have to define coefficients for each level of the factor we want to compare. To compare Extended daylight and Flashlights, subtract one effect from the other, as shown in the second row of Fig. 8.7. To compare the control (O) with the other two, you subtract the effect of O from the mean of the effects of the two treatments; see the first row of Fig. 8.7. GenStat will now use these coefficients to split up the treatment sum of squares.

Technically, you have filled the matrix so that the sum of the coefficients is zero for each comparison, i.e. for each row in Fig. 8.7. In Fig. 8.7, this is the case for each comparison: (-1) + 0.5 + 0.5 = 0 and –1 + 1 = 0. Also the sum of the pairwise products of the coefficients is zero: (-1)*0+0.5*(-1)+0.5*1 =0. This is the definition of orthogonal contrasts. Orthogonal contrasts can be interpreted separately because they are estimated independently, though useful contrasts do not have to be orthogonal.

Fig. 8.8 The contrasts are incorporated in the treatment structure



Design: General Analysis of Variance

Y-Variate: Eggs

Treatment Structure: COMP(Treatment;2;Conf)

Block Structure: Block

Interactions: All Interactions

☐ Covariates

When the matrix is ready, click somewhere outside the matrix to send the data to the GenStat server and to return to the main dialogue. You will see (Fig. 8.8) that the treatment structure in the Analysis of Variance dialogue window has changed. Click on ***[OK]*** and the resulting Treatment sum of squares in the ANOVA table in the Output Window has been split up. The Treatment effects explain about half (4212.5) of the total variation (8864.2). Almost all of those Treatment effects are explained by the difference between the control and the two lighting regimes (4134.4 of 4212.5), while the difference between extended daylight and flashing explains almost nothing (78.1 of 4212.5). So the significant difference between treatments we saw before is due to the difference between O on the one hand and E and F on the other hand (p-value of 0.017). There is no evidence for a difference between E and F.

```
***** Analysis of variance *****

Variate: Eggs

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              3     2330.3      776.8    2.01

Block.*Units* stratum
Treatment                  2     4212.5     2106.3    5.44  0.045
  O vs E and F             1     4134.4     4134.4   10.69  0.017
  E vs F                   1       78.1       78.1    0.20  0.669
Residual                   6     2321.5      386.9

Total                     11     8864.2
```

No multiple comparison tests ?

If you have been using other statistical packages to analyse experimental data, then you may have been using multiple comparison tests (Newman-Keuls, Tukey, Duncan, etc.) to compare treatments, rather than the contrasts that we described above.

We are pleased to report that the GenStat Discovery Edition menus do NOT include multiple comparison tests. The developers of GenStat do not feel that these add to the proper analysis of experimental data. They did introduce these tests in the latest version so that they could demonstrate that these tests are not of value!

So you will have to pay for the latest version of GenStat if you want to do these tests. You can also see in the guides of SSC Reading why we do feel they are not worthwhile.

8.3 Defining the treatment structure

We have analysed two designs so far: a completely randomised design (or a one-way ANOVA without blocking) and a one-way ANOVA in randomised blocks. In each case there has been one single factor identifying the individual treatments. In the ANOVA dialogue window, when you click on the arrow next to the design-box, you can see a whole range of other designs that can be analysed.

Designs in ANOVA menu:

One-way ANOVA (no Blocking)

One-way ANOVA (in Randomized Blocks)

Two-way ANOVA (no Blocking)

Two-way ANOVA (in Randomized Blocks)

Completely Randomized Design

Split-Plot Design

Split-split Plot Design

Latin Square

Graeco-Latin Square

Lattice Design

You can also choose ***General Analysis of Variance*** and we suggest you use this option routinely. Once you understand a few basics you will be able to specify the correct analysis for more complex designs.

The treatment and layout (blocking) structure are entered using a formula in which following operators are used:

+ addition	e.g. A+B+C main effects of A, B, and C
. interaction	e.g. A.B interaction of A and B
* cross-product	A*B is equivalent to A+B+A.B
/ nesting	A/B is equivalent to A+A.B

We explore this for two examples of a factorial treatment structure, one in a randomised block layout, and the other in split plots.

8.3.1 Factorial treatment structure

Factorial experiments are studies during which direct effects of more than one experimental treatment are examined simultaneously, while at the same time cross-effects or interactions of those treatments are examined. The advantages of factorial experiments are summarised as

- if there are no interactions, you benefit from “hidden replication”
- if there are interactions, the trial can investigate them

Let’s start with a theoretical example. An experiment is set up to examine the influence of variety, the application of an insecticide and the application of a fungicide on the yield of maize. Let’s investigate some possible approaches.

Treatment	Description
A	Variety 1, no insecticide, no fungicide
B	Variety 2, no insecticide, no fungicide
C	Variety 1, insecticide, no fungicide
D	Variety 2, no insecticide, fungicide

If the experiment were set up as in the table above, there would be some problems caused by the way the experiment was designed. When comparing treatments using a one–way ANOVA, only the differences between treatment A and B will be caused by the variety effect. The differences between C and D could be due to variety, insecticide or fungicide and the effect of variety is measured only for no insecticide or fungicide.

An alternative designe, with 8 treatments, is

Treatment	Description
A	Variety 1, no insecticide, no fungicide
B	Variety 2, no insecticide, no fungicide
C	Variety 1, insecticide, no fungicide
D	Variety 2, insecticide, no fungicide
E	Variety 1, insecticide, fungicide
F	Variety 2, insecticide, fungicide
G	Variety 1, no insecticide, fungicide
H	Variety 2, no insecticide, fungicide

If the experiment were set up as in this table, you could analyse using a one-way ANOVA and calculate some confidence intervals as seen above. The effect of the variety could be found by examining the difference between treatments A and C and E and G versus B and D and F and H.

This ANOVA is completely valid, but the results are meaningful **only if** you are willing to accept that the effect of each factor is the same at all levels of the other factors: changes in maize yield due to the use of a different variety are the same if an insecticide is applied or not **and** if a fungicide is applied or not **AND** changes in maize yield due to the use of an insecticide are the same for both varieties **and** if a fungicide is applied or not **AND** changes in maize yield due to the use of a fungicide are the same for both varieties **and** if an insecticide is applied or not. So, this ANOVA is valid under the assumption of the treatment effects being **additive** (we used the word “and” quite often in the previous sentence). There is however no way that we can prove that this assumption is true and if the assumption is not true, the results are only valid for each specific set of levels that were compared. For instance the Variety effect would only be valid if no insecticide and no fungicide are applied.

Although there are situations where it makes no sense to look at interactions, most of the time it does and a factorial treatment structure will give much more information for this type of analyses.

The design of the theoretical example should be rewritten as a design with 3 factors in which all the levels of each factor are combined with each other. This factorial treatment structure would look as follows.

(Treatment number)	Variety	Insecticide	Fungicide
1	V1	No	No
2	V2	No	No
3	V1	Yes	No
4	V2	Yes	No
5	V1	No	Yes
6	V2	No	Yes
7	V1	Yes	Yes
8	V2	Yes	Yes

Now we can find the main effects of each factor, the average change for different levels of the other factors, by examining following differences:

- Main effect of Variety = treatments 1, 3, 5, 7 versus 2, 4, 6, 8
- Main effect of Insecticide = treatments 1, 2, 5, 6 versus 3, 4, 7, 8
- Main effect of Fungicide = treatments 1, 2, 3, 4 versus 5, 6, 7, 8

We can also investigate interactions. Does using an insecticide has for instance the same effect on both varieties? In this example, this would be examined by examining treatments 1 and 5 versus 2 and 6 (the effect of variety when no insecticide is applied) and by examining treatments 3 and 7 versus 4 and 8 (the effect of variety when an insecticide is applied).

GenStat doesn't have a special option to analyse three-way ANOVA's. Instead, we use the General Analysis of Variance design. In the ***Treatment structure*** box, we type a formula using the operators mentioned in the introduction above. For the example this would be (Fig. 8.9):

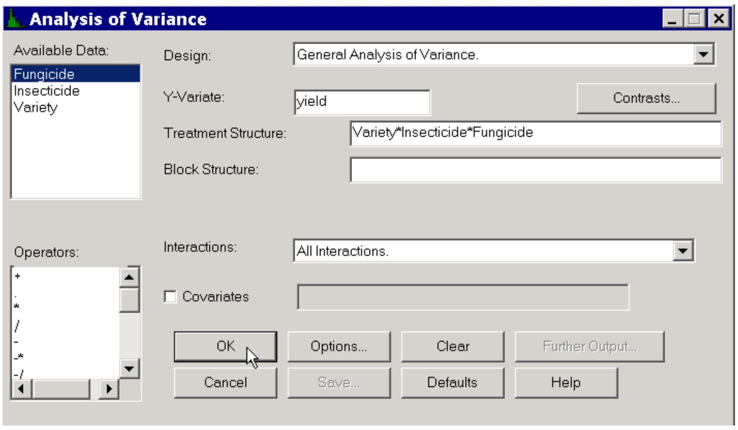
Variety*Insecticide*Fungicide

This can be rewritten as:

Variety+Insecticide+Fungicide+Variety.Insecticide+Variety.Fungicide+
Insecticide.Fungicide+Variety.Insecticide.Fungicide

So, a combination of the main effect of each factor, the first order interactions and the second order interactions.

Fig. 8.9 An example of a factorial treatment structure



Analysis of Variance

Available Data: Fungicide, Insecticide, Variety

Design: General Analysis of Variance.

Y-Variate: yield

Treatment Structure: Variety*Insecticide*Fungicide

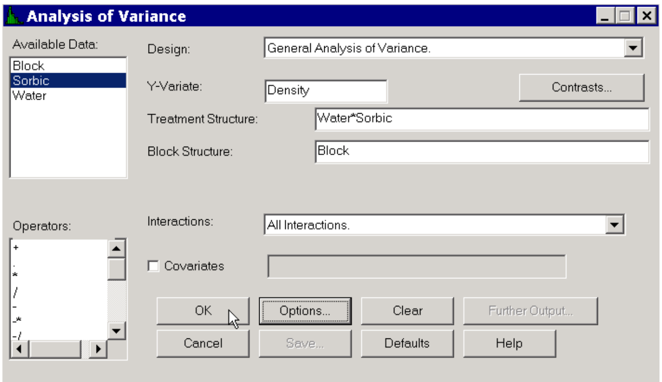
Block Structure:

Interactions: All Interactions.

☐ Covariates

We now analyse the dataset “*Salmonella typhirum survival.xls*” that was considered in chapter 6. Open this file. We want to analyse the variate *Density*, the design is factorial with two factors *Water* and *Sorbic* and the experimental units are grouped in randomised blocks. Complete the Analysis of Variance dialogue window as shown in Fig. 8.10 and indicate in the ANOVA Options window that you also want the LSDs to appear in the Output Window.

Fig. 8.10 Factorial treatment structure



Available Data: Block, Sorbic, Water

Design: General Analysis of Variance.

Y-Variate: Density

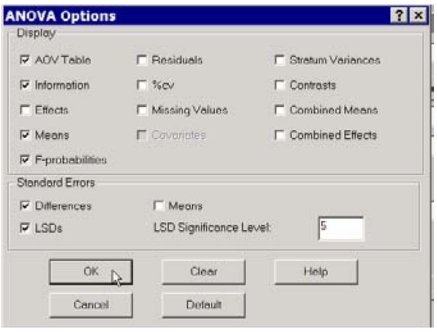
Treatment Structure: Water*Sorbic

Block Structure: Block

Interactions: All Interactions.

☐ Covariates

Fig. 8.11 Indicating you want the LSDs



Display

☑ AOV Table ☐ Residuals ☐ Stratum Variances

☑ Information ☐ %cv ☐ Contrasts

☐ Effects ☐ Missing Values ☐ Combined Means

☑ Means ☐ Covariates ☐ Combined Effects

☑ F-probabilities

Standard Errors

☑ Differences ☐ Means

☑ LSDs LSD Significance Level: 5

The Output includes the ANOVA table, with sums of squares for each factor and also for the interaction.

There are three tables of means in the output, corresponding to each of the terms in the treatment formula, one each for the main effects of *Sorbic* and *Water*, and one for the full treatment set (interaction table).

The tables of standard errors of differences of means and LSDs each have three columns to match the three tables of means. For example in the ‘Least significant differences’ table the value of 0.136 given under ‘*Sorbic*’ is the LSD for comparing Sorbic means – these are means of 18 (= 3 blocks x 6 levels of Water) data values (‘rep’) and the LSD (and equivalent SED) is based on 34 d.f. (The Residual d.f.).

```
***** Analysis of variance *****

Variate: Density

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              2    0.01385    0.00692    0.17

Block.*Units* stratum
Water                      5   81.56910   16.31382  403.72  <.001
Sorbic                     2    2.75936    1.37968   34.14  <.001
Water.Sorbic              10    1.31626    0.13163    3.26  0.005
Residual                  34    1.37389    0.04041

Total                     53   87.03245

 * MESSAGE: the following units have large residuals.
 Block 3     *units* 9           0.41   s.e. 0.16
 Block 3     *units* 10         -0.42   s.e. 0.16


***** Tables of means *****

Variate: Density

Grand mean  5.50

    Water     0.78     0.82     0.86     0.90     0.94     0.98
              4.19     4.39     4.79     5.53     6.42     7.68

   Sorbic    0 ppm  100 ppm  200 ppm
              5.80     5.44     5.26

    Water   Sorbic    0 ppm  100 ppm  200 ppm
     0.78              4.28     4.23     4.07
     0.82              4.56     4.30     4.30
     0.86              5.14     4.70     4.53
     0.90              6.00     5.36     5.24
     0.94              6.53     6.41     6.31
     0.98              8.30     7.67     7.08

 *** Standard errors of differences of means ***

Table                    Water      Sorbic       Water
                                                Sorbic
rep.                         9          18           3
d.f.                        34          34          34
s.e.d.                   0.095       0.067       0.164

*** Least significant differences of means (5% level) ***

Table                    Water      Sorbic       Water
                                                Sorbic
rep.                         9          18           3
d.f.                        34          34          34
l.s.d.                   0.193       0.136       0.334
```

The GenStat commands that were produced to execute the analysis can be seen in the input log (***Window ⇒ Input log***).

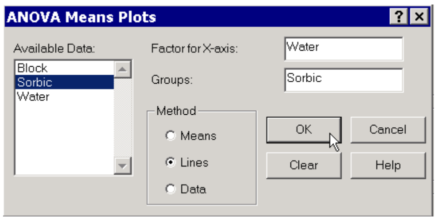
```
"General Analysis of Variance."
BLOCK Block
TREATMENTS Water*Sorbic
COVARIATE "No Covariate"
ANOVA [PRINT=aovtable,information,means; FACT=32; FPROB=yes;
PSE=diff,lsd; LSDLEVEL=5]\
 Density
```

When interpreting the results from a factorial analysis of variance with an interaction, it is often useful to give a pictorial representation of the 2-way table of means. One method of doing this within GenStat is to select the ***[Further Output]*** button in the ***Analysis of Variance*** dialogue box, then click ***[Means Plots]***.

> This plot allows the table of means to be plotted against one of the factors. If an analysis of variance has been carried out, with two treatment factors in a factorial combination, then one of the factors is chosen as the ***Factor for X-axis***. The means will be plotted against this factor. The other factor defines the ***Groups***. The means for each level of the ***Groups*** factor will be distinguished with different colours and/or symbols. By default, just the means are plotted, but if ***Lines*** is selected under ***Method*** instead, then the means will be connected by lines.

In this example choose *water* to be the factor whose levels are on the x-axis, and *sorbic* to be the groups factor, as shown in Fig. 8.12. Select Lines as the Method of plotting. The resulting graph is shown in Fig. 8.13.

Fig. 8.12 Selecting the factor for the X-axis and the factor for the groups



ANOVA Means Plots

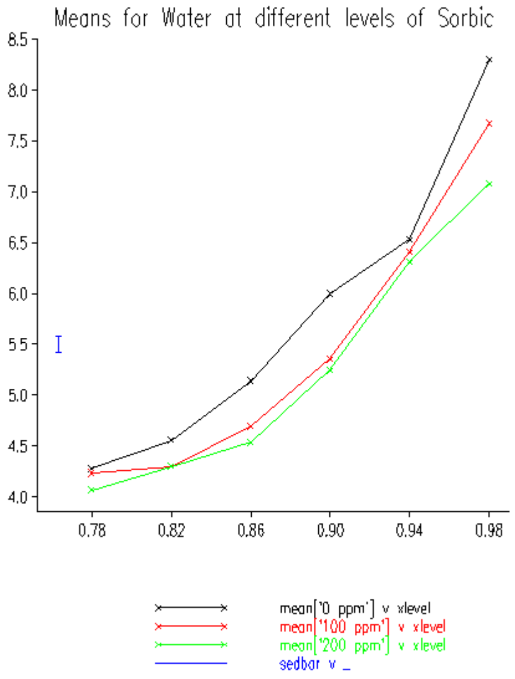
Available Data:
Block
Sorbic
Water

Factor for X-axis: Water

Groups: Sorbic

Method
○ Means
● Lines
○ Data

Fig. 8.13 The resulting graph



The graph includes an SED bar, which is centred about the grand mean.

It is sometimes more helpful to give the LSD bar. As yet, this cannot be done directly with the menu, but this can be achieved by modifying the command the menu produces. The modified line can then be executed with ***Run ⇒ Line***. ***LSDs*** was selected under ***[Options]*** of the Analysis of Variance dialogue box, so the appropriate LSD value has already been calculated and printed (0.334).

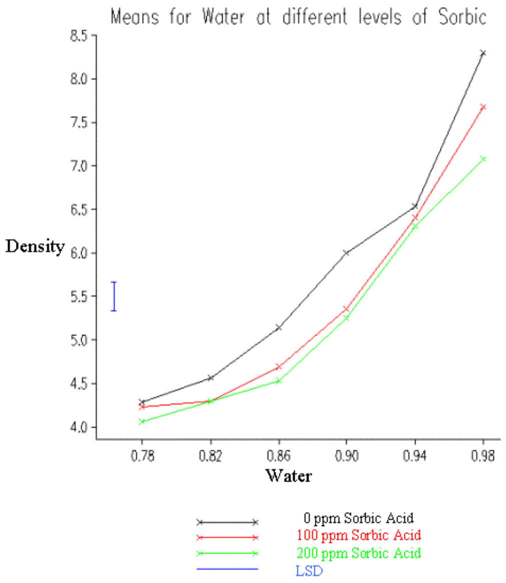
The use of the dialogue above, followed by the editing of the line in the Input log has sent the single line **AGRAPH [method=lines] Water;Sorbic** to the GenStat server. Once you become more experienced, a much quicker alternative way is to type the command directly. This is best done in a new Input window. Alternatively you can modify commands from the Input log or in the Output Window. If you add "; **bar=0.334** " to the end of this line (as below),

AGRAPH [method=lines] Water;Sorbic;bar=0.334

Leave the cursor in the line and choose ***Run ⇒ Submit Line***, to give a graph with the LSD bar as shown in Fig. 8.14

If you add that command to those that are given in the Input log earlier (see page 94) and submit everything to the GenStat server, GenStat will carry out the ANOVA and plot the graph automatically.

Fig. 8.14 The same graph but with an LSD bar



Refer to chapter 6.3 for adding textboxes to the graph. Use ***Run ⇒ Restart Session*** to clear the data.

8.3.2 Nested blocking structure

In some factorial experiments, it is necessary for practical reasons to use larger experimental units for some factors than for others. This is called a Split Plot design. For a Split Plot experiment with the main plots laid out in a Randomised Block design, a factor is needed for blocks (*block*), another for main plots within blocks (*mainplot* – to which levels of treatment *factor1* are applied) and another for subplots within main plots (*subplot* – to which levels of treatment *factor2* are applied). In general, there can be more than one factor at either level.

The general formula describing the layout and the factorial treatments of a split-plot design are :

Layout : block/mainplot/subplot

Treatment factors : factor1*factor2

The operator "/" is the nesting operator. For instance the formula

A/B

would expand to A+A.B The latter term can be thought of as "B within A".

Enter this in the General Analysis of Variance dialogue box or alternatively, select Split-plot from the list of possible designs. The three layout factors then have to be entered in the ***Blocks, Whole Plots***, and ***Sub-plots*** boxes.

Both dialogues generates the following GenStat commands:

```
BLOCK block/mainplot/subplot
TREATMENTS factor1*factor2
```

The BLOCK statement can be translated as 'subplots nested within mainplots which are nested within blocks'. GenStat is excellent for analysing designs with this greater level of complexity, as all the information required is given by the formulae entered into the boxes defining the layout (***Blocks, Whole Plots*** etc) and, separately, the ***Treatment Structure*** (or with the equivalent BLOCK and TREATMENTS statements). The output is comprehensive as all the standard errors of differences are calculated, even for multi-way treatment tables, along with the degrees of freedom required for each.

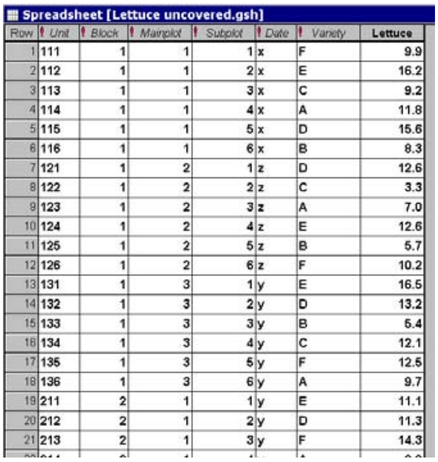
Example: from Mead, Curnow and Hasted, pages 151-155 (Example 7.4). Six varieties of lettuce are grown in frames and the frames are removed on several dates. There are 4 blocks (1, 2, 3, 4), each with 3 main plots for the 3 different uncovering dates (x, y, z). Within each main plot there are 6 subplots for the 6 varieties (A, B, C, D, E, F). The data (yield of lettuce) were originally given in systematic order. For the purposes of illustrating the factors required in GenStat, these data are presented below as from a possible experiment.

Block 1							Block 2							Block 3							Block 4						
111	1	1	1	x	F	9.9	211	2	1	1	Y	E	11.1	311	3	1	1	x	F	4.8	411	4	1	1	z	D	11.3
112			2		E	16.2	222			2		D	11.3	312			2		D	10.3	412			2		A	6.3
113			3		C	9.2	223			3		F	14.3	313			3		C	11.4	413			3		F	12.2
114			4		A	11.8	224			4		A	8.8	314			4		B	11.8	414			4		B	8.8
115			5		D	15.6	225			5		B	12.9	315			5		A	9.7	415			5		C	2.6
116			6		B	8.3	226			6		C	15.7	316			6		E	14.0	416			5		E	14.1
121	1	2	1	z	D	12.6	221	2	2	1	Z	F	11.6	321	3	2	1	y	B	11.2	421	4	2	1	x	F	9.8
122			2		C	3.3	222			2		B	8.4	322			2		D	11.0	422			2		B	8.5
123			3		A	7.0	223			3		A	9.1	323			3		F	15.9	423			3		C	7.2
124			4		E	12.6	224			4		E	12.3	324			4		C	7.6	424			4		D	14.7
125			5		B	5.7	225			5		C	6.9	325			5		E	10.8	425			5		A	6.4
126			6		F	10.2	226			6		D	15.4	326			6		A	12.5	426			6		E	11.5
131	1	3	1	y	E	16.5	231	2	3	1	X	C	10.6	331	3	3	1	z	E	14.4	431	4	3	1	y	E	8.5
132			2		D	13.2	232			2		B	8.4	332			2		A	7.1	432			2		F	7.5
133			3		B	5.4	233			3		A	7.5	333			3		C	1.0	433			3		C	9.4
134			4		C	12.1	234			4		F	10.8	334			4		D	14.2	434			4		A	9.4
135			5		F	12.5	235			5		D	10.8	335			5		F	10.4	435			5		B	7.8
136			6		A	9.7	236			6		E	11.2	336			6		B	6.1	436			6		D	10.7
Unit	Block	Main Plot	Sub-plot	Date	Variety	Yield	Unit	Block	Main Plot	Sub-plot	Date	Variety	Yield	Unit	Block	Main Plot	Sub-plot	Date	Variety	Yield	Unit	Block	Main Plot	Sub-plot	Date	Variety	Yield

In the plan above, the first four entries (columns) in each plot are the unit, block, mainplot and subplot numbers (blocking factors). These layout factors are systematically ordered within the design. They are followed by the labels for the treatment factors for date and variety, together with the yields that are to be analysed.

Enter the data into a spreadsheet (e.g. column by column from the plan above), with three factors (*block, mainplot*, and *subplot*) which define the layout of the experiment (these will be in a systematic order). Two factors are created to indicate the treatment factors *date* and *variety*. The spreadsheet is shown in Fig. 8.15. Save the data, giving the file the name “L*ettuce uncovered.gsh*”

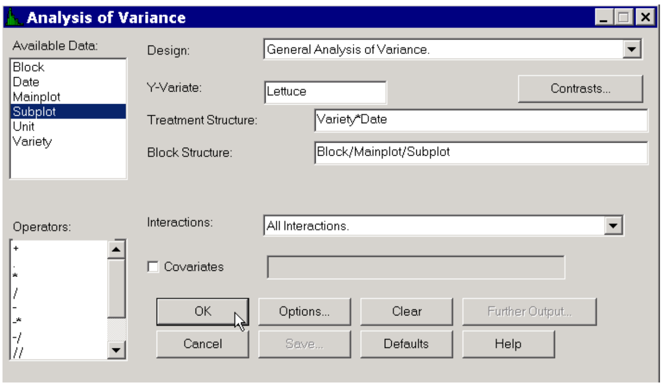
Fig. 8.15 The spreadsheet with the lettuce data



Spreadsheet [Lettuce uncovered.gsh]

Row	Unit	Block	Mainplot	Subplot	Date	Variety	Lettuce
1	111	1	1	1	x	F	9.9
2	112	1	1	2	x	E	16.2
3	113	1	1	3	x	C	9.2
4	114	1	1	4	x	A	11.8
5	115	1	1	5	x	D	15.6
6	116	1	1	6	x	B	8.3
7	121	1	2	1	z	D	12.6
8	122	1	2	2	z	C	3.3
9	123	1	2	3	z	A	7.0
10	124	1	2	4	z	E	12.6
11	125	1	2	5	z	B	5.7
12	126	1	2	6	z	F	10.2
13	131	1	3	1	y	E	16.5
14	132	1	3	2	y	D	13.2
15	133	1	3	3	y	B	5.4
16	134	1	3	4	y	C	12.1
17	135	1	3	5	y	F	12.5
18	136	1	3	6	y	A	9.7
19	211	2	1	1	y	E	11.1
20	212	2	1	2	y	D	11.3
21	213	2	1	3	y	F	14.3

Fig. 8.16 Factorial treatment structure and nested blocking structure



Analysis of Variance

Available Data: Block, Date, Mainplot, Subplot, Unit, Variety

Design: General Analysis of Variance.

Y-Variate: Lettuce

Treatment Structure: Variety*Date

Block Structure: Block/Mainplot/Subplot

Interactions: All Interactions.

□ Covariates

Operators: + . * / - -* -/ //

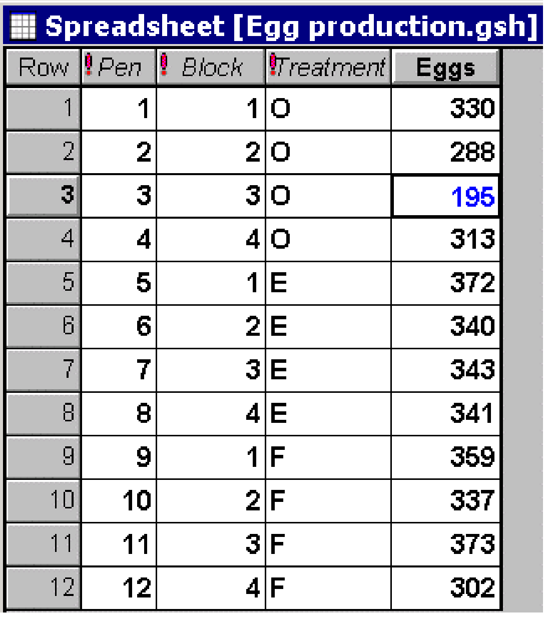
In the Analysis of Variance dialogue box, choose the General Analysis of Variance and complete as in Fig. 8.16. Notice that the two SEDs required for the variety by date table of means are printed in the right-hand column of the Standard errors table. The second (under the section headed ‘Except when comparing means with the same level(s) of’) is for comparing two means with the same *date* (1.59). The first SED for the *variety* by *date* table (1.65) is to compare two means for different dates. The relevant d.f. are printed immediately below each SED. Use ***Run ⇒ Restart Session*** to clear the data.

```
***** Analysis of variance *****

Variate: Lettuce
Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              3     29.343      9.781    1.35

Block.Mainplot stratum
Date                       2     38.003     19.002    2.62  0.152
Residual                   6     43.566      7.261    1.44

Block.Mainplot.Subplot stratum
Variety                    5    260.508     52.102   10.32  <.001
Variety.Date              10    163.698     16.370    3.24  0.003
Residual                  45    227.277      5.051

Total                     71    762.395

* MESSAGE: the following units have large residuals.
Block 1     Mainplot 3     Subplot 1            4.3   s.e. 1.8
Block 1     Mainplot 3     Subplot 3           -4.4   s.e. 1.8

***** Tables of means *****
Variate: Lettuce

Grand mean  10.3

  Variety        A        B        C        D        E        F
               8.8      8.6      8.1     12.6     12.8     10.8

     Date        x        y        z
              10.4     11.1      9.3

  Variety     Date        x        y        z
        A               8.9     10.1      7.4
        B               9.3      9.3      7.2
        C               9.6     11.2      3.4
        D              12.8     11.5     13.4
        E              13.2     11.7     13.3
        F               8.8     12.6     11.1

*** Standard errors of differences of means ***
Table              Variety        Date     Variety
                                              Date
rep.                    12          24           4
s.e.d.                0.92        0.78        1.65
d.f.                    45           6       46.05
Except when comparing means with the same level(s) of
 Date                                         1.59
 d.f.                                           45
```

8.3.3 Checking for outliers

The nesting approach can also be used to increase the information shown in the Output Window. Start a new session and open the "*Egg production.gsh*" file again. Each unit is a pen of 6 chickens. We can consider each unit as a level of a factor "*Pen*". So, insert a column containing this factor before the column "*Block*" as in Fig. 8.17. Now we will add a mistake to the dataset to be sure it will contain an outlier. Change the number of eggs laid by Pen 3 (Block 3, Treatment O) from 295 to 195.

Fig. 8.17 Adding a unique identifier for each unit



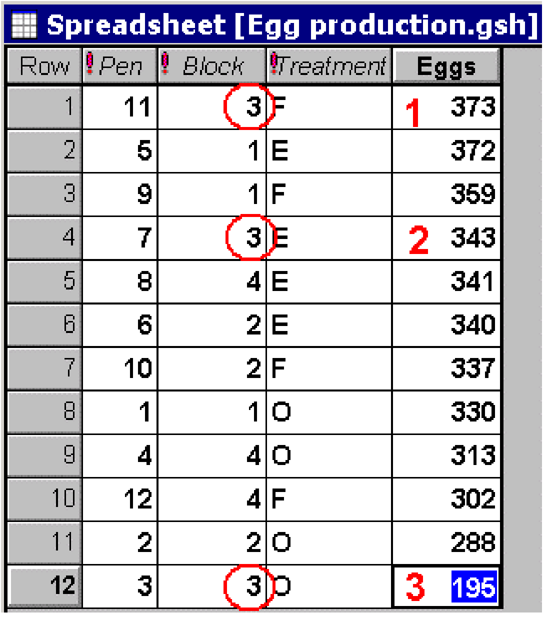
Spreadsheet [Egg production.gsh]

Row	Pen	Block	Treatment	Eggs
1	1	1	O	330
2	2	2	O	288
3	3	3	O	195
4	4	4	O	313
5	5	1	E	372
6	6	2	E	340
7	7	3	E	343
8	8	4	E	341
9	9	1	F	359
10	10	2	F	337
11	11	3	F	373
12	12	4	F	302

```
***** Analysis of variance *****

Variate: Eggs

Source of variation       d.f.        s.s.        m.s.     v.r.   F pr.

Block stratum                3       3980.       1327.     0.76

Block.*Units* stratum
Treatment                    2      11129.       5565.     3.19   0.114
Residual                     6      10472.       1745.

Total                       11      25581.


* MESSAGE: the following units have large residuals.

Block 3      *units* 1            -66.    s.e. 30.
```

When you carry out the ANOVA, you will get a warning that one observation has large standardized residuals. But you have to start counting to find this observation; in this case it is the first unit of block 3 (Fig. 8.18). Now, this is not very convenient when you analyse large and complex datasets. Furthermore, the message will be different when the order of data changes.

Fig. 8.18 Counting to find the observation with a large standardized residual



Spreadsheet [Egg production.gsh]

Row	Pen	Block	Treatment	Eggs
1	11	3	F	373
2	5	1	E	372
3	9	1	F	359
4	7	3	E	343
5	8	4	E	341
6	6	2	E	340
7	10	2	F	337
8	1	1	O	330
9	4	4	O	313
10	12	4	F	302
11	2	2	O	288
12	3	3	O	195

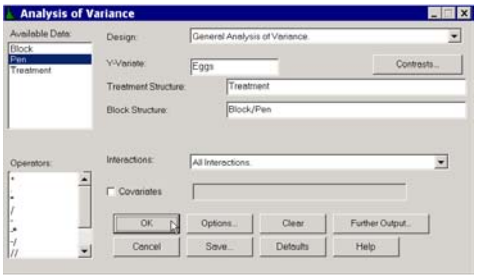
For instance, the same ANOVA on the same dataset but this time ordered according to the descending number of eggs would result in following warning.

```
***** Analysis of variance *****

Variate: Eggs

Source of variation      d.f.        s.s.        m.s.     v.r.  F pr.

Block stratum               3       3980.       1327.     0.76

Block.*Units* stratum
Treatment                   2      11129.       5565.     3.19  0.114
Residual                    6      10472.       1745.

Total                      11      25581.


* MESSAGE: the following units have large residuals.

Block 3      *units* 3             -66.    s.e. 30.
```

However, the unit number, in this case the pen number, could be considered as a layout factor that is nested within the blocks. So we could also carry out an ANOVA as in Fig. 8.19

Fig. 8.19 Putting the units in the block structure

```
"General Analysis of Variance."
BLOCK Block/Pen
TREATMENTS Treatment
COVARIATE "No Covariate"
ANOVA
[PRINT=aovtable,information,means;
FACT=32; FPROB=yes; PSE=diff] Eggs
```



Analysis of Variance

Available Data: Block, Pen, Treatment

Design: General Analysis of Variance

Y-Variate: Eggs

Treatment Structure: Treatment

Block Structure: Block/Pen

Interactions: All Interactions

□ Covariates

The resulting error message will now give very direct information regardless of how the data were ordered, in this example that the observation with the large residual is observation Pen 4 from Block 3.

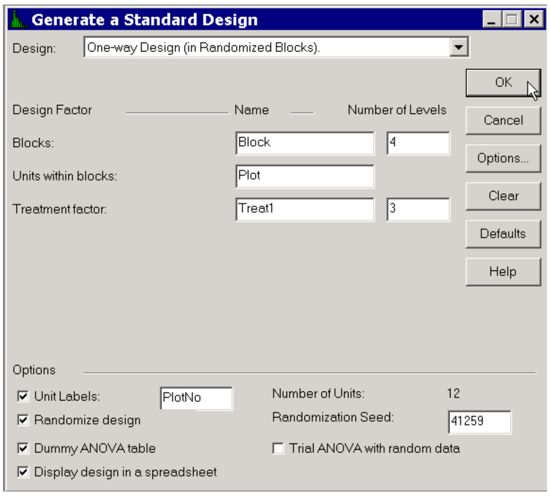
```
***** Analysis of variance *****

Variate: Eggs

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              3      3980.      1327.    0.76

Block.Pen stratum
Treatment                  2     11129.      5565.    3.19  0.114
Residual                   6     10472.      1745.

Total                     11     25581.

* MESSAGE: the following units have large residuals.

Block 3     Pen 4             -66.   s.e. 30.
```

8.4 Randomising an Experiment

GenStat includes the facilities for the randomisation of a wide range of designs. For those that are used in this guide the procedure is to use ***Stats ⇒ Design ⇒ Generate a Standard Design.***

Here we first show how to design a randomised block experiment of the same size as analysed earlier, namely with 4 blocks and 3 treatments. Select ***One-way Design (in Randomized blocks).*** Complete the dialogue as shown in Fig. 8.20, by inserting the information for the blocks and treatments. All the rest can remain as the defaults, so click on ***[OK]***. The results are as shown in the spreadsheet in Fig. 8.21.

Fig. 8.20 The dialogue window for generating a standard design



Generate a Standard Design

Design: One-way Design (in Randomized Blocks).

Design Factor	Name	Number of Levels
Blocks:	Block	4
Units within blocks:	Plot	
Treatment factor:	Treat1	3

Options

☑ Unit Labels: PlotNo

☑ Randomize design

☑ Dummy ANOVA table

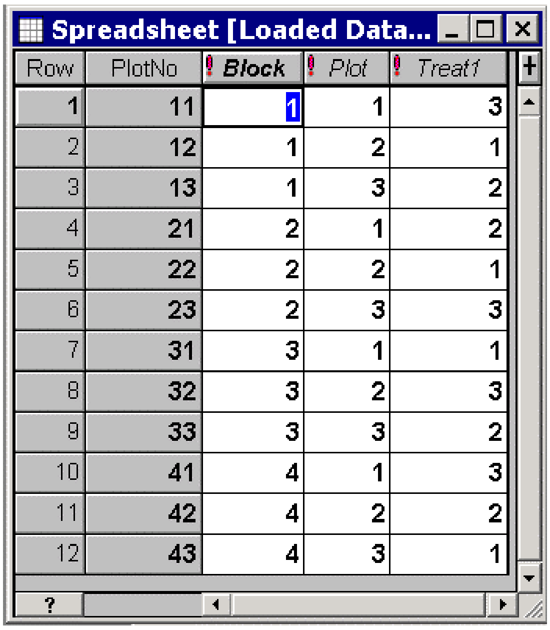
☑ Display design in a spreadsheet

Number of Units: 12

Randomization Seed: 41259

☐ Trial ANOVA with random data

Fig. 8.21 The resulting spreadsheet



Spreadsheet [Loaded Data...

Row	PlotNo	Block	Plot	Treat1
1	11	1	1	3
2	12	1	2	1
3	13	1	3	2
4	21	2	1	2
5	22	2	2	1
6	23	2	3	3
7	31	3	1	1
8	32	3	2	3
9	33	3	3	2
10	41	4	1	3
11	42	4	2	2
12	43	4	3	1

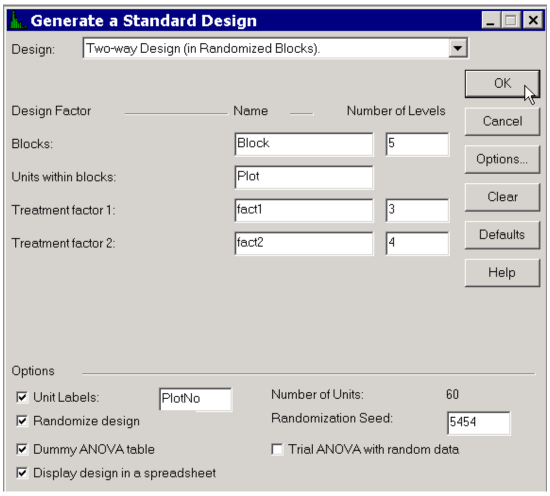
Note that GenStat has added a column to give the plot number, as well as one, called *Plot*, which gives the plot number within each block. The order of the values in your treatment column will probably not be the same as above, but you could get the identical randomisation if you use the same number for the Seed, as is shown in the dialogue box above. This '*Randomization Seed*' is an initial value GenStat needs for a subroutine that generates a random numbers.

The design information has been copied into a spreadsheet. This can now be saved as a GenStat spreadsheet, if you intend later to enter the experimental data directly into GenStat. Alternatively it can be saved in a standard spreadsheet format, such as an Excel file, if you will use that for the design of data collection forms and then for the data entry.

The next example is an experiment with a factorial treatment structure. Use ***Run ⇒ Restart Session*** to clear the data and then ***Stats ⇒ Design ⇒ Generate a Standard Design***. As an example we take a Randomised Block design with 5 blocks of 12 treatment combinations comprising the factorial set for two factors, one with 3 levels (*fact1*) and the other with 4 levels (*fact2*).

Complete the menu as shown in Fig. 8.22. It gives the plan in a spreadsheet (Fig. 8.23) and also the dummy ANOVA table in the output window, as shown below.

Fig. 8.22 Generating a two-way design in randomised blocks



Generate a Standard Design

Design: Two-way Design (in Randomized Blocks).

Design Factor	Name	Number of Levels
Blocks:	Block	5
Units within blocks:	Plot	
Treatment factor 1:	fact1	3
Treatment factor 2:	fact2	4

Options

☑ Unit Labels: PlotNo

☑ Randomize design

☑ Dummy ANOVA table

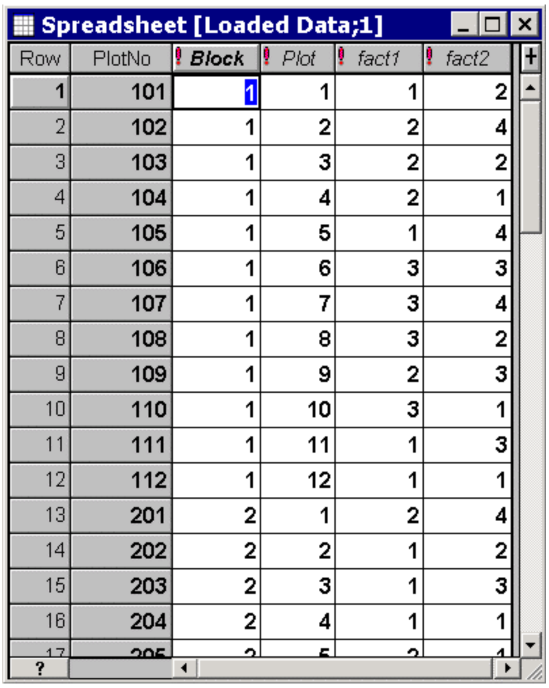
☑ Display design in a spreadsheet

Number of Units: 60

Randomization Seed: 5454

☐ Trial ANOVA with random data

Fig. 8.23 The resulting spreadsheet



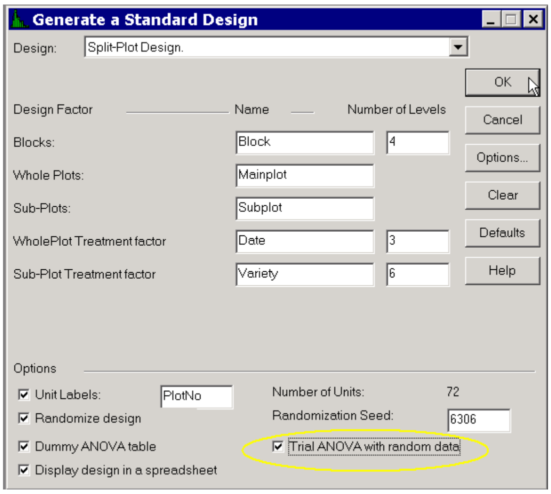
Spreadsheet [Loaded Data;1]

Row	PlotNo	Block	Plot	fact1	fact2
1	101	1	1	1	2
2	102	1	2	2	4
3	103	1	3	2	2
4	104	1	4	2	1
5	105	1	5	1	4
6	106	1	6	3	3
7	107	1	7	3	4
8	108	1	8	3	2
9	109	1	9	2	3
10	110	1	10	3	1
11	111	1	11	1	3
12	112	1	12	1	1
13	201	2	1	2	4
14	202	2	2	1	2
15	203	2	3	1	3
16	204	2	4	1	1
17	205	2	5	2	4

```
***** Analysis of variance *****
Source of variation     d.f.

Block stratum              4

Block.Plot stratum
fact1                      2
fact2                      3
fact1.fact2                6
Residual                  44

Total                     59
```

The final example shows the completion of the same dialogue for a split plot design, with the same structure as was analysed earlier in this guide. Note when completing the dialogue a further option is to request an ANOVA with trial data. This illustrates the form that the results will be presented in.

Fig. 8.24 Generating a split-plot design



Design: Split-Plot Design.

Design Factor	Name	Number of Levels
Blocks:	Block	4
Whole Plots:	Mainplot	
Sub-Plots:	Subplot	
WholePlot Treatment factor	Date	3
Sub-Plot Treatment factor	Variety	6

Options

☑ Unit Labels: PlotNo

☑ Randomize design

☑ Dummy ANOVA table

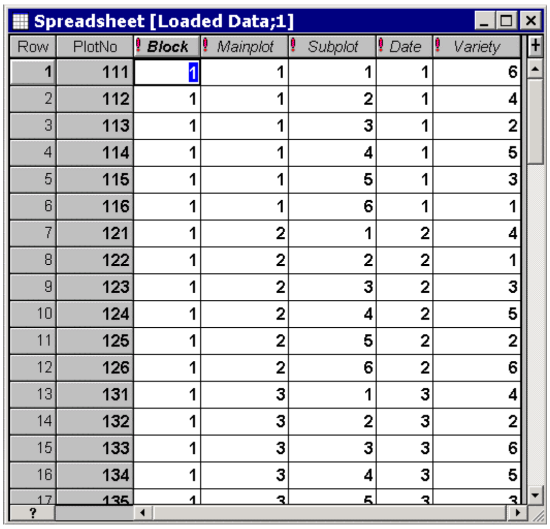
☑ Display design in a spreadsheet

Number of Units: 72

Randomization Seed: 6306

☑ Trial ANOVA with random data

Fig. 8.25 The resulting spreadsheet



Row	PlotNo	Block	Mainplot	Subplot	Date	Variety
1	111	1	1	1	1	6
2	112	1	1	2	1	4
3	113	1	1	3	1	2
4	114	1	1	4	1	5
5	115	1	1	5	1	3
6	116	1	1	6	1	1
7	121	1	2	1	2	4
8	122	1	2	2	2	1
9	123	1	2	3	2	3
10	124	1	2	4	2	5
11	125	1	2	5	2	2
12	126	1	2	6	2	6
13	131	1	3	1	3	4
14	132	1	3	2	3	2
15	133	1	3	3	3	6
16	134	1	3	4	3	5
17	135	1	3	5	3	3

```
***** Analysis of variance *****

Variate: _Rand_

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              3    215.770     71.923   13.15

Block.Mainplot stratum
Date                       2      0.865      0.433    0.08  0.925
Residual                   6     32.807      5.468    5.47

Block.Mainplot.Subplot stratum
Variety                    5      1.981      0.396    0.40  0.849
Date.Variety              10      5.563      0.556    0.56  0.840
Residual                  45     45.000      1.000

Total                     71    301.987

***** Tables of means *****

Variate: _Rand_

Grand mean  13.28

     Date        1        2        3
             13.43    13.20    13.21

  Variety        1        2        3        4        5        6
             12.95    13.36    13.29    13.42    13.21    13.44

Date  Variety     1        2        3        4        5        6
  1           13.58    13.27    13.11    13.47    13.43    13.74
  2           12.88    13.43    12.86    13.55    13.16    13.32
  3           12.39    13.37    13.92    13.25    13.04    13.26


*** Standard errors of means ***

Table                 Date     Variety        Date
                                           Variety
rep.                    24          12           4
e.s.e.               0.477       0.289       0.660
d.f.                     6          45       19.78
Except when comparing means with the same level(s) of
 Date                                        0.500
 d.f.                                           45

*** Least significant differences of means (5% level) ***

Table                 Date     Variety        Date
                                           Variety
rep.                    24          12           4
l.s.d.               1.652       0.822       1.950
d.f.                     6          45       19.78
Except when comparing means with the same level(s) of
 Date                                        1.424
 d.f.                                           45


***** Stratum standard errors and coefficients of variation *****
Variate: _Rand_

Stratum                     d.f.         s.e.         cv%

Block                          3        1.999        15.1
Block.Mainplot                 6        0.955         7.2
Block.Mainplot.Subplot        45        1.000         7.5
```

9 Challenge 3

The design of the experiment in Challenge 1 (page 57) was a randomised block design, with the blocks specified in column REP. Use the analysis of variance to find the standard error of the difference between Sesbania (code SES) and natural fallows (code NAT) in (a) mean maize yield (b) soil nitrate.

10 Further Reading

10.1 Other free documentation

A second part of this guide, called "Further regression and ANOVA using GenStat Discovery Edition", is being produced. As the title says, it includes further information on regression and on the analysis of variance. It will be distributed on future versions of the GenStat Discovery Edition CD and will also be available from the GenStat for Africa website: www.worldagroforestrycentre.org/genstatforafrica

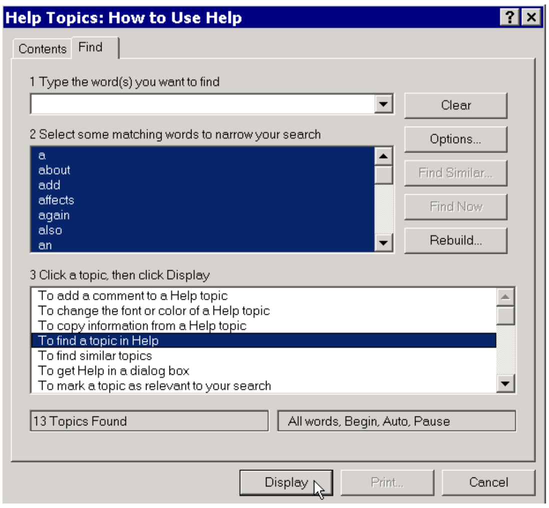
ICRAF, the World Agroforestry Centre, has prepared a set of notes on the analysis of experimental data. These include analyses using GenStat, plus a set of sample data files. The materials are on the CD and are also available from the website of ICRAF's Research Support Unit: www.worldagroforestrycentre.org/rsu (under Data Analysis). On the same website, under Resources, you find several technical notes. Some of them contain additional information on GenStat.

The Statistical Services Centre of the University of Reading (http://www.rdg.ac.uk/ssc/) , has produced a series of "good-practice" guides. They are on the CD and are also available on the SSC website: http://www.rdg.ac.uk/ssc/develop/dfid/booklets.html They include a guide to help MSTAT users who would like to start using GenStat. The Biometry Unit Consultancy Services (BUCS) of the University of Nairobi, in collaboration with statisticians from Malawi and Zimbabwe, has produced a guide on their strategy for statistical software in their agriculture faculties. They propose the use of GenStat for postgraduate students and research. For undergraduates they suggest the use of SSC-Stat (an add-in for Excel) and Instat+ (a simple statistics package). In addition to GenStat Discovery Edition, the CD also includes both SSC-Stat and Instat. Up-to-date versions of these packages may be downloaded from the SSC website. Some presentations on the BUCS strategies can be found at http://www.uonbi.ac.ke/acad_depts/bucs/presentation.htm The whole BUCS website is also available on the CD.

10.2 The Help menu

We showed in chapter 2.2.2 (page 10) how to find more information on a subject using the GenStat Help. In the example, we saw how to learn more about the different spreadsheet formats that can be imported into GenStat. The GenStat help looks and works rather similar to many other Windows programs, but if in doubt on how to use it, browse first through the "how to use help" menu. Choose ***Help => How to use help*** and select for instance "*To find a topic in Help*". Click on the ***[Display]*** button as in Fig. 10.1 and a help window will pop up containing information on finding a topic in the GenStat Help (Fig. 10.2).

Fig. 10.1 Learning more about the GenStat Help



Help Topics: How to Use Help

1 Type the word(s) you want to find

2 Select some matching words to narrow your search

a
about
add
affects
again
also
an

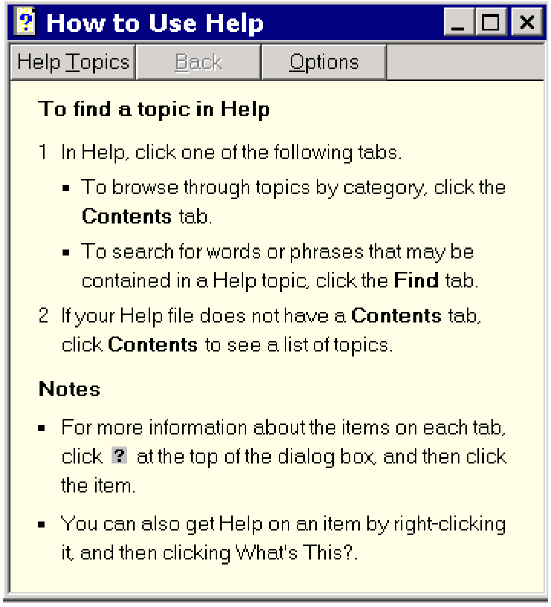
3 Click a topic, then click Display

To add a comment to a Help topic
To change the font or color of a Help topic
To copy information from a Help topic
To find a topic in Help
To find similar topics
To get Help in a dialog box
To mark a topic as relevant to your search

13 Topics Found

All words, Begin, Auto, Pause

Fig. 10.2 The resulting information on how to find a topic in the GenStat Help



How to Use Help

To find a topic in Help

1 In Help, click one of the following tabs.

- To browse through topics by category, click the **Contents** tab.
- To search for words or phrases that may be contained in a Help topic, click the **Find** tab.

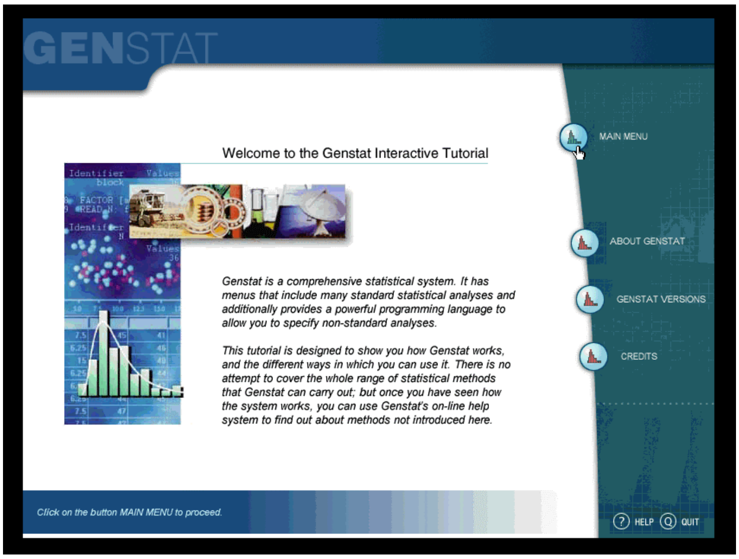
2 If your Help file does not have a **Contents** tab, click **Contents** to see a list of topics.

Notes

- For more information about the items on each tab, click ? at the top of the dialog box, and then click the item.
- You can also get Help on an item by right-clicking it, and then clicking What's This?.

Choosing ***Help=> GenStat Tutorial*** opens a tutorial covering similar subjects as in this guide but in an interactive way: partly text, partly video, partly interactive pop-up windows. Click on the ***[Main Menu]*** button (Fig. 10.3) to start the tutorial.

Fig. 10.3 Starting the GenStat Tutorial



Welcome to the Genstat Interactive Tutorial

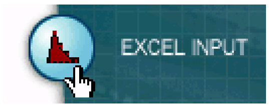
Genstat is a comprehensive statistical system. It has menus that include many standard statistical analyses and additionally provides a powerful programming language to allow you to specify non-standard analyses.

This tutorial is designed to show you how Genstat works, and the different ways in which you can use it. There is no attempt to cover the whole range of statistical methods that Genstat can carry out; but once you have seen how the system works, you can use Genstat's on-line help system to find out about methods not introduced here.

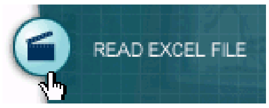
Click on the button MAIN MENU to proceed.

You navigate through the tutorial by clicking on different types of buttons (Fig. 10.4).

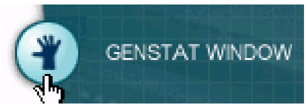
Fig. 10.4 Different types of buttons in the GenStat Tutorial and their meaning



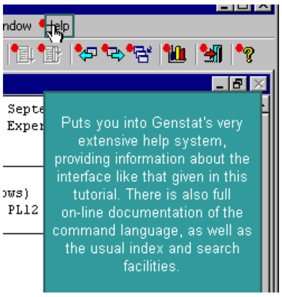
go to a specific section



start a movie clip about the topic



open an interactive page about the topic



Puts you into Genstat's very extensive help system, providing information about the interface like that given in this tutorial. There is also full on-line documentation of the command language, as well as the usual index and search facilities.

On such an interactive page, move the cursor over the red dots and a window will pop up containing more information, in this case about the different menu options.

You leave the tutorial by clicking on a ***[Quit]*** button, or first by clicking on a ***[Back]*** button until you see a ***[Quit]*** button, and confirming that you want to quit by clicking ***[Yes]***.

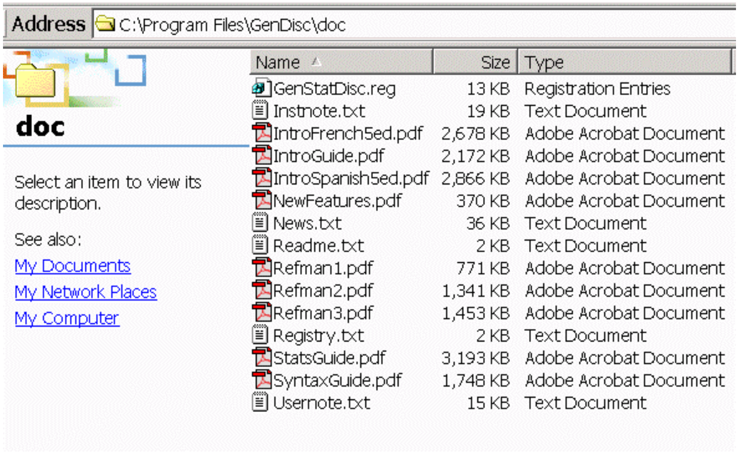
10.3 “Hidden” user guides

And there is more! After installing GenStat Discovery edition, you have somewhere hidden on your computer almost 3,000 pages of user guides in pdf format. Pdf format stands for portable data format and is a file format that can be read with Adobe Acrobat Reader. This software is free and is likely to be on your machine. If not, a copy is included on the CD. You can get the latest version at www.adobe.com

The reason for the fact that those files are not easily visible is that the guides are intended to be included in the Help menu of GenStat version 6. VSN International, the producers of GenStat, decided shortly before the release of the Discovery Edition to also make these guides available for Discovery Edition users but there was no time left to change the menus. Although the guides are intended for version 6, most of the information is still useful for the GenStat Discovery Edition (which is basically version 5). Only the sections on graphics are likely to differ.

If you installed GenStat Discovery Edition in a standard way, you find the pdf-files containing the documentation in the subfolder: C:\Program Files\GenDisc\doc (Fig. 10.5).

Fig. 10.5 The folder containing additional information



Address C:\Program Files\GenDisc\doc

doc

Select an item to view its description.

See also:

My Documents

My Network Places

My Computer

Name	Size	Type
GenStatDisc.reg	13 KB	Registration Entries
Instnote.txt	19 KB	Text Document
IntroFrench5ed.pdf	2,678 KB	Adobe Acrobat Document
IntroGuide.pdf	2,172 KB	Adobe Acrobat Document
IntroSpanish5ed.pdf	2,866 KB	Adobe Acrobat Document
NewFeatures.pdf	370 KB	Adobe Acrobat Document
News.txt	36 KB	Text Document
Readme.txt	2 KB	Text Document
Refman1.pdf	771 KB	Adobe Acrobat Document
Refman2.pdf	1,341 KB	Adobe Acrobat Document
Refman3.pdf	1,453 KB	Adobe Acrobat Document
Registry.txt	2 KB	Text Document
StatsGuide.pdf	3,193 KB	Adobe Acrobat Document
SyntaxGuide.pdf	1,748 KB	Adobe Acrobat Document
Usernote.txt	15 KB	Text Document

Following table gives an overview of the different documents.

Introguide.pdf	Roger Payne, Darren Murray, Simon Harding, David Baird, Duncan Soutar & Peter Lane. 2002. GenStat® for WindowsTM (6th Edition) Introduction. VSN International, Oxford, UK. 276 pp. ISBN-1-904375-06-5
NewFeatures.pdf	Roger Payne (Ed.) 2002. New features in GenStat® Release 6.1 VSN International, Oxford, UK. 95 pp. ISBN 1-904375-02-2
SyntaxGuide.pdf	Roger Payne (Ed.). 2002. The Guide to GenStat® Release 6.1 Part 1: Syntax and Data Management. VSN International, Oxford, UK. 492 pp. ISBN 1-904375-00-6
StatsGuide.pdf	Roger Payne (Ed.). 2002. The Guide to GenStat® Release 6.1 Part 2: Statistics VSN International, Oxford, UK. 856 pp. ISBN 1-904375-01-4
Refman1.pdf	Roger Payne et al. 2002. GenStat® Release 6.1 Reference Manual Part 1: Summary. VSN International, Oxford, UK. 254 pp. ISBN 1-904375-03-0
Refman2.pdf	Roger Payne et al. 2002. GenStat® Release 6.1 Reference Manual Part 2: Directives VSN International, Oxford, UK. 396 pp. ISBN 1-904375-04-9
Refman3.pdf	Roger Payne and Gillian Arnold (Eds.) 2002. GenStat® Release 6.1 Reference Manual Part 3: Procedure Library PL14 VSN International, Oxford, UK. 454 pp. ISBN 1-904375-05-7

10.4 Non-English speakers

In the same directory, you find two introductory guides on using GenStat for Windows version 5, translated in French and Spanish. GenStat Discovery Edition is exactly the same as version 5 except for the graphics.

IntroFrench5ed.pdf	Simon Harding, Peter Lane, Darren Murray et Roger Payne. Traduit par Gaston Kokodé. 2000. Genstat pour Windows (5éme Edition) Introduction VSN International sarl, Oxford, UK. 216 pp. ISBN 1-85206-183-9
IntroSpanish5ed.pdf	Simon Harding, Peter Lane, Darren Murray y Roger Payne. Traducido al español por Guillermo Hough y Freddy Ledezma. 2000. Genstat para Windows (5ta. Edición) Introducción VSN Internacional Ltda., Oxford, UK. 216 pp. ISBN 1-85206-183-9

Also, the guide that you are reading now is being translated in French and might be on the CD already. It will be available on the GenStat for Africa website: www.worldagroforestrycentre.org/genstatforafrica

10.5 The GenStat user community

Finally, there is an informal community of GenStat users who are active trough a mailing list. You can read the rules of the list, browse the archives and join the list at http://www.bioss.sari.ac.uk/genstat/