Introduction to
Instat+

Roger Stern, Joan Knock, Colin Grayer
and Sandro Leidi

Draft

June 2002
# Contents

**Acknowledgements** ........................................................................................................ 1

**Part 1 – Getting Started** .................................................................................. 3

**Chapter 1 - Introduction** .................................................................................. 5

1.1 The context ........................................................................................................ 5
1.2 Who might use Instat .......................................................................................... 5
1.3 History .................................................................................................................. 6
1.4 Other statistical packages .................................................................................. 6

**Chapter 2 - Getting Started** ............................................................................. 7

2.1 Introduction .......................................................................................................... 7
2.2 Getting and installing Instat ................................................................................ 7
2.3 Using an Instat tutorial ....................................................................................... 8
2.4 Using this introductory guide .............................................................................. 9
2.5 Registering Instat .............................................................................................. 11
2.6 Technical details of Instat’s installation ............................................................. 12

**Chapter 3 - Data sets** ......................................................................................... 15

3.1 Introduction ......................................................................................................... 15
3.2 A regression example - (File: regress.wor) ......................................................... 16
3.3 A survey - (File: survey.wor) ............................................................................. 17
3.4 An experiment - (File: experi.wor) ................................................................... 19
3.5 Monthly rainfall records - (File: rain.wor) ......................................................... 21
3.6 Data at multiple levels (File: activity.wor) ......................................................... 22
3.7 Poverty data (Files: poverty.wor and poverty.sav) ......................................... 23
3.8 An ecological study (Files: coral.xls and coral.wor) ......................................... 24

**Chapter 4 - Using a statistics package** ............................................................... 27

4.1 Introduction ......................................................................................................... 27
4.2 Using menus ...................................................................................................... 29
4.3 Dialogues ............................................................................................................ 30
4.4 Windows ............................................................................................................. 33
4.5 More than menus and dialogues ....................................................................... 36
4.6 Making mistakes ............................................................................................... 38
4.7 Starting with Excel ............................................................................................ 39
4.8 In conclusion ..................................................................................................... 41

**Chapter 5 - Getting help** .................................................................................. 45

5.1 Introduction ......................................................................................................... 45
5.2 Help on using Instat .......................................................................................... 45
5.3 Help on statistics .............................................................................................. 49
5.4 Help on teaching or learning statistics .............................................................. 50
5.5 In conclusion ..................................................................................................... 52

**Part 2 – Organising the data** ............................................................................. 55

**Chapter 6 - Worksheets** .................................................................................. 57

6.1 Introduction ......................................................................................................... 57
6.2 Creating a new worksheet ................................................................................. 57
6.3 Using an existing Instat worksheet .................................................................. 58
6.4 Using Instat’s worksheet editor ....................................................................... 59
6.5 The data in a worksheet ................................................................................... 59
6.6 Navigating round the worksheet ..................................................................... 62
6.7 The information about a worksheet ................................................................. 63
6.8 In conclusion ..................................................................................................... 64

28/06/2002
<table>
<thead>
<tr>
<th>Chapter 13 - Tables</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1 Introduction</td>
<td>139</td>
</tr>
<tr>
<td>13.2 Frequency tables</td>
<td>140</td>
</tr>
<tr>
<td>13.3 Summary tables</td>
<td>142</td>
</tr>
<tr>
<td>13.4 Presenting tables effectively</td>
<td>145</td>
</tr>
<tr>
<td>13.5 Teaching percentages and other topics</td>
<td>148</td>
</tr>
<tr>
<td>13.6 Weights in tables</td>
<td>151</td>
</tr>
<tr>
<td>13.7 General tables</td>
<td>152</td>
</tr>
<tr>
<td>13.8 Multiple responses</td>
<td>154</td>
</tr>
<tr>
<td>13.9 Data at multiple levels</td>
<td>157</td>
</tr>
<tr>
<td>13.10 Missing values</td>
<td>160</td>
</tr>
<tr>
<td>13.11 Saving the information in the table</td>
<td>160</td>
</tr>
<tr>
<td>13.12 In conclusion</td>
<td>164</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 14 - Probability models</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.1 Introduction</td>
<td>167</td>
</tr>
<tr>
<td>14.2 The role of modelling and inference</td>
<td>167</td>
</tr>
<tr>
<td>14.3 Looking at the normal distribution</td>
<td>168</td>
</tr>
<tr>
<td>14.4 Looking at other distributions</td>
<td>170</td>
</tr>
<tr>
<td>14.5 Calculating probabilities and percentage points</td>
<td>171</td>
</tr>
<tr>
<td>14.6 In conclusion</td>
<td>173</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 15 - Fitting simple models</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.1 Introduction</td>
<td>175</td>
</tr>
<tr>
<td>15.2 Confidence intervals for data from a normal distribution</td>
<td>175</td>
</tr>
<tr>
<td>15.3 Comparing 2 samples from normal distributions</td>
<td>178</td>
</tr>
<tr>
<td>15.4 Looking at proportions – an example of a non-normal model</td>
<td>180</td>
</tr>
<tr>
<td>15.5 Simple regression – emphasising the general concepts</td>
<td>182</td>
</tr>
<tr>
<td>15.6 In conclusion</td>
<td>186</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 4 – Further Methods</th>
<th>Page</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Chapter 16 - Analysis of variance</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.1 Introduction</td>
<td>189</td>
</tr>
<tr>
<td>16.2 One-way ANOVA</td>
<td>190</td>
</tr>
<tr>
<td>16.3 Including blocks in the model</td>
<td>192</td>
</tr>
<tr>
<td>16.4 Simple use of the orthogonal ANOVA dialogue</td>
<td>193</td>
</tr>
<tr>
<td>16.6 Factorial treatment structure</td>
<td>196</td>
</tr>
<tr>
<td>16.7 A split plot experiment</td>
<td>199</td>
</tr>
<tr>
<td>16.8 Examining the treatment means</td>
<td>201</td>
</tr>
<tr>
<td>16.9 Coping with common problems</td>
<td>206</td>
</tr>
<tr>
<td>16.10 In conclusion</td>
<td>207</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 17 - Regression models</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.1 Introduction</td>
<td>211</td>
</tr>
<tr>
<td>17.2 Correlations</td>
<td>212</td>
</tr>
<tr>
<td>17.3 Simple linear regression</td>
<td>213</td>
</tr>
<tr>
<td>17.4 Adding structure – comparison of regressions</td>
<td>214</td>
</tr>
<tr>
<td>17.5 Multiple regression</td>
<td>217</td>
</tr>
<tr>
<td>17.6 General linear models</td>
<td>219</td>
</tr>
<tr>
<td>17.7 Regression diagnostics</td>
<td>221</td>
</tr>
<tr>
<td>17.8 Lack of Fit</td>
<td>226</td>
</tr>
<tr>
<td>17.9 Polynomial and periodic regression models</td>
<td>228</td>
</tr>
<tr>
<td>17.10 Weights and missing values in regression</td>
<td>232</td>
</tr>
<tr>
<td>17.11 In conclusion</td>
<td>232</td>
</tr>
</tbody>
</table>
Acknowledgements

Over the years, many people have contributed to Instat, either by writing modules, testing the package or helping with the documentation.

We would like to particularly mention:

First, Bob Burn, who started the whole saga of Instat by writing some statistics programs for the BBC microcomputer in the early 1980s. Bob wrote the first version of the Reference Guide and the program module for regression analysis.

Joy Merritt, Lesley Francis, Philip Swandale, Jackie Tyler and Sue Miller, who wrote some of the modules for Instat. David Stern and Sachin Medakkar wrote most of the new tables module.

Richard Russell for producing BBCBASIC(86), for writing some routines in assembler code and for advising us on aspects of the PC version of Instat.

Clifford Pearce, who wrote the original GenAnova program and allowed us to adapt and include the program in Instat.

David Baird, who adapted his DataLoad and ODBC programs for Instat.

The Overseas Branch of the UK Met Office, and particularly Steve Palmer, who funded part of the development of Instat into Instat+.

Richard Coe, Eleanor Allan, Carlos Barahona, Mark Woodward and staff and students of the Department of Applied Statistics and the Statistical Services Centre at The University of Reading.

Alex Owen, who typed and retyped this guide as we changed our mind, and still managed to keep her sense of humour.

Thanks are also due to the staff and participants on training courses at Reading and elsewhere, plus many who have used the earlier versions of Instat and who have helped us with suggestions and encouraging remarks.
Part 1 – Getting Started

The first part of this guide consists of Chapters 1 to 5. It serves to introduce Instat and statistics packages in general.

In Chapters 1 and 2 we describe who might find Instat useful and how to install the software.

The whole guide is written in “tutorial” fashion, so users can try any of the examples that are illustrated. We use a wide variety of datasets, and several are described in Chapter 3.

In Chapter 4 we consider how to use a statistics package. Our description is more general than Instat and other statistics packages are introduced in the last section of the chapter.

Statistical packages here become a useful resource for supporting information and, in Chapter 5, we describe the Help in Instat and in other statistical software.
Chapter 1 - Introduction

1.1 The context

Instat is a simple general-purpose statistics package. It is designed to support the teaching of statistics, while being powerful enough to assist research in any discipline that requires the statistical analysis of data.

Instat is also part of a “resource pack” of materials that can support trainers who would like to change the way they teach statistics. Other parts of this pack, developed at Reading and elsewhere, include an “add-in” for Excel and a series of “good-practice” guides. These guides and other components are described in Chapter 5 and some can be accessed from within Instat’s help system. The Excel add-in is described in Chapter 4.

1.2 Who might use Instat

One use of Instat is for teachers of statistics who would like to include more practical illustrations of the concepts in their course.

Other users are those who now need to analyse data but they found that their statistics courses did not prepare them either for the volume of data or the methods they need to use, once they start their research. This is sometimes because courses are too “method-based”. We complement that approach by making this guide more “data-based”. We suggest courses for non-statisticians could add components on the organisation of data, described here in Part 2 of this guide and devote more time to the descriptive methods described in Chapters 11 to 13.

If you need to understand the basic concepts of statistics, then Instat and this guide may help. Perhaps you have taken courses in statistics, but still do not feel confident that you understand the main ideas. They are described in Chapters 14 and 15.

If you currently just use a spreadsheet for your data analysis, and wonder whether this is sufficient, then this guide and Instat may help you to define your strategies. We do not suggest that you abandon your spreadsheet. Far from it. But using more than one package is now easy, and perhaps a statistics package can be used for those parts of the analysis that are more difficult with a spreadsheet. We suggest a spreadsheet is sufficient up to the topics covered in Chapter 15.

Instat will just be a stepping-stone for some spreadsheet users. A few years ago, the major statistics packages were challenging to use. But this has changed, and those who find that Instat adds to their skills in data analysis will then find it is also easy to graduate to a more advanced package, if their needs outpace Instat’s facilities. What they have learned with Instat should also help them to choose which of the many statistics packages they will find to be the most appropriate.

Perhaps you just need a free statistics package. Instat is free for all individual users.

If you currently use a statistics package, for example Minitab, or SPSS or Genstat, then adding Instat may help you to use it better. Do not stop using that package, because Instat is not as powerful as the commercial statistics packages.

In some of our short training courses we now show examples using more than one statistics package and offer a range of packages, including Instat, for practical work. These are courses that are designed to teach statistics, rather than statistical computing. It is now so easy to switch between Windows’ based packages that some participants find that they are better able to understand the concepts if they use the most appropriate package for each part of the course.
Similarly, for undergraduate and postgraduate students, courses used to be linked to a particular statistics package. We now permit students to use whichever package they wish, because our aim is to teach the statistical ideas and not merely the package itself. Adding Instat is an easy option, and students who wish to use it on their own machine can download it.

1.3 History

Other types of software seem quite recent, by comparison with statistics packages. For example spreadsheets started with VisiCalc, and then Lotus 123 was popular until the mid 1990’s. Now many people use Excel. In word processing the mid 1980’s saw Wordstar as the standard, then it was WordPerfect, and now the leader is Word.

Most statistics packages have a much longer history. The two giants are SAS and SPSS and they both started in the 1960’s. If you think that computers only started in the 1940’s and personal computers did not appear until the 1980’s, then this is a long time ago!

Instat started in 1983, as a set of programs to help the analysis of experimental data for a training course in Sri Lanka. It was initially written for the BBC microcomputer that was popular in the UK at that time. Later Instat graduated to the ordinary PC, and added components particularly for processing climatic data.

Until 2000, Instat was just a “DOS-based” package. Then we were able to add a Windows front-end and this is the hybrid that is currently available. We still use the old “DOS-based back-end” to do most of the processing, but have written a new “front-end” in Visual Basic, so you never need to use the backend yourself. We have also added a spreadsheet view of the data and a text editor.

We have started to implement components that are independent of the past. In particular the new module on tables, described in Chapter 13, does not use the backend at all.

1.4 Other statistical packages

There are many general statistics packages. We view Instat as an introductory statistics package. For some, it may be enough, but others might use it as a stepping-stone and later move to, or add another statistics package.

The last section of most chapters in this guide is called “in conclusion”. There we mainly describe facilities that are available in other software and extend what has been described for Instat within the chapter. We have avoided any comparison of these packages and simply give examples of the good practice that these packages promote. In giving these examples we have no “hidden agenda” to promote any particular package, nor should the omission of any package be taken to mean that it does not also promote good practice for the aspects that are described.

What we are trying to promote is the idea that an efficient strategy for data processing is likely to involve users, or at least organisations, in having multiple packages available. This is recent, because it follows from the ease of transfer of data between packages, the simplicity of use of modern software, and the similarity of the interface that the different packages offer to the user. We return to these issues in Chapter 10 of this guide.
Chapter 2 - Getting Started

2.1 Introduction

In this chapter, we describe how to install Instat in Section 2.2. Once installed, it is useful to distinguish between the two following tasks.

- Learning how to use Instat.
- Using Instat to learn about particular sets of data.

We believe it is important to spend an initial period learning about the software, using one or two of the data sets that we have provided. One way is with one of the Instat tutorials that are in printed form and on-line, as part of the Help. We describe that route in Section 2.3. A second possibility is to use some of the materials from this guide and we give suggestions in Section 2.4.

Users who are familiar with other statistics packages may find they need only about an hour for this introduction to the software. Others may need half a day or more.

Instat is designed to be accessible to “beginners”, including those who have always found statistics to be a difficult subject. Indeed one of our aims is to help such users overcome some of their fear of statistics.

By a “beginner” we mean a beginner to the use of a computer to statistics, not a beginner to computers. If you are new to Windows, then we find that users can not concentrate on statistics, or Instat, or any other statistics package. They continually have to practice basic Windows, or mouse skills, or where to find files. If you are really new to computers, then we strongly suggest that you separate learning basic computing skills from your introduction to a statistics package.

2.2 Getting and installing Instat

Instat is available on CD, or it may be downloaded from the web site of the Statistical Services Centre at The University of Reading, UK. The web site is on http://www.rdg.ac.uk/ssc. Either way, you could look occasionally on the web site for information, upgrades and for information on registering Instat. We describe registration in Section 2.5.

On individual machines Instat installs as any other Windows package. By default it goes into the C: \ Program Files \ directory.

Instat can be run over a network. In this case it is first installed and tested on the network server. Then individual machines need to say where the central copy is located, so the installation copies the files that are needed to each machine. This is often done remotely by the staff who administer the network. We give technical details of the installation in the last section of this Chapter. They are mainly intended for those who are installing Instat for multiple use.

Instat is not large as a program, but we do include a lot of additional material. You need about 20 Mbytes for the software and supporting materials, such as the Instat Guides.
2.3 Using an Instat tutorial

Two tutorials are supplied with Instat. The first is general and we expect that most new users will start there. The second is a shorter version, designed mainly for those who will use Instat mainly for the analysis of climatic data.

These tutorials are available as part of the Help in Instat, as shown in Fig. 2.3a. At some stage you may also be tempted by the “Getting Started in Instat” option, part of which is shown in Fig. 2.3b.

The contents, if you use the “Tutorial Guide” option, are included in Fig. 2.3c. If you use this guide on-line, then it is designed to fill just half the screen, as shown in Fig. 2.3d so you can organise Instat to fill the other half. Then use the option within the Help to keep it “on top”.

The tutorial guides are also available as printed guides and in pdf form, so you can print your own copy.
2.4 Using this introductory guide

If you are reasonably experienced in the use of the computer for statistical work, then you may not need the Tutorial Guide, or just reading it may be enough.

This Introductory Guide has been designed so that you can dip into the different chapters. You might want to look briefly at Chapter 3, where we describe some of the data sets used for illustration throughout this guide. Then Chapter 4 describes different ways of using Instat and Chapter 5 looks at getting Help.

The second part of this guide, Chapters 6 to 10, is called “Organising the data”. It is concerned mainly with the topics from the Manage menu, shown in Fig. 2.4a. We find that many analysis are less effective than they could be, because users have not organised their data sufficiently. Look at some of these chapters, if such issues are of concern, or leave them until later.

Within this second part of the guide, Chapter 10 is more general and discusses possible strategies for the use of statistical software.
Part 3 is called “Basic Statistics”. This includes descriptive statistics, which is covered in Chapters 11 to 13. The Chapter 11 looks at graphical summaries, and Chapter 12 at numerical summaries and one-way tables. Chapter 13 covers tabulation in general, and particularly for survey data.

The concepts of basic statistical inference, are in the next two chapters, with Chapter 14 devoted to probability models and Chapter 15 on simple inference. Within Instat’s menus, shown in Fig. 2.4b, these five chapters have mainly described the use of Graphics, and Simple Models.

**Fig. 2.4b** Instat with the statistics menu

The last part of the guide is titled “Further Statistical Methods”. Two key chapters are on the Analysis of Variance, Chapter 16, and Regression, Chapter 17.

The following chapters consider the remaining methods listed in the menu shown in Fig. 2.4b. In Chapter 18 we look at the processing of qualitative data. We describe the general ideas and then look at non-parametric methods and also at log-linear models.

In Chapter 19 we consider the ways in which data can be processed when “time” is important. This includes survival studies, where the time to an event has been recorded, as well as standard time series analyses. In Chapter 20 we look at the generation of random samples. In Instat this is mainly to support teaching, but the general subject of simulation is also mentioned.

The final substantial chapter looks at the use of Instat to support the design phase of a research study. There we also explain why it is the last main chapter!

Throughout this guide we use Instat’s menus and dialogues. But Instat can also be used through typing commands, and we explain in Chapter 4 why this is sometimes a good idea. In Appendix 1 we explain more about the command mode of Instat, and also how simple macros can be written.
2.5 Registering Instat

Instat may be distributed freely and used for private, non-commercial work at no charge. Until you register there is a (slightly annoying) nag screen to tempt you to contact us for a registration number. Registration is free for non-commercial use. We do not pass your contact details to any third party; your registration is useful for us to have an indication of who may be using Instat and where they work. The form, which is available on our web site, www.rdg.ac.uk/ssc/ is shown in Fig. 2.5a. If you do not have good access to the web, you are welcome to send details by e-mail to instat@reading.ac.uk or to write to us.

Fig. 2.5a Registering for Instat on the web

If you download Instat/Instat+ from this website, we would be grateful if you would send us a few details about yourself. By filling in the form below, you are helping us to understand more about where and how Instat is being used. Your details will be added to our database so that we can - if you agree - send you occasional news on Instat+ related matters.

Your name: 
Your e-mail address: 
Subscribe to the mailing list? ☐ Yes please ☐ No thanks 
Organisation: 
Postal Address: 
Country: 
Please mention briefly your area of work: 
Any comments or suggestions:

Once you register, we may contact you very occasionally by e-mail to tell you of further developments of Instat or other SSC activities. This will be rare, because we assume most people prefer to keep up-to-date by looking at our web site occasionally. If you do not want further messages, then just tell us.

There are reasonable licence fees for multiple use of Instat within a non-commercial environment and for single and multiple use commercially. More details of the different licences are on our web site.
2.6 Technical details of Instat’s installation

This section is for those who would like to understand precisely where Instat puts files, when it is installed. It is particularly for users installing Instat over a network, and is not needed by someone who just wants to use Instat on a single-user machine.

We have followed the guidelines in the “Applications Specification for Microsoft® Windows® 2000” available from http://msdn.microsoft.com/certification/appspec.asp. These rules are intended particularly for Windows® 2000 or later operating systems, but they make good sense also for Windows® 95, 98, NT and ME systems.

The program is installed by default within the C:\Program Files\Instat folder, though the destination folder is an option within the installation process. There are then three sub-folders for the program, the help files and the library of datasets. The macro libraries are in further subfolders of the program folder. In a network installation, these folders are on the server and, in general, users do not have write access to these folders.

We need a transfer folder to pass information between the Instat server and the front end. This is always stored locally in the folder designated for this purpose by the system. This is often the C:\Windows\Temp folder. We also have two configuration files, called insdata.bbc and insdata.dat within which we store configuration information. The master copy that we supply is within the program folder and a copy is made into the user’s application data folder, the first time that Instat is run. This is often the folder C:\Windows\Application data\, within which we make a subfolder called SSC, and a further subfolder called Instat within this.

When you first run Instat you will be asked for a folder within which you wish to work. The default is usually the “My Documents” folder. We suggest that you use subfolders of this folder, but you are welcome to choose any other folder, as long as it is one to which you have write access.

A temporary copy of the configuration files is put into the working folder. The menu Edit ⇒ Options, shown in Fig. 2.6a may be used at any time to change your configuration options. These include the working directory, default worksheet sizes and so on. When changes are made, they may be for just the current session, in which case only the temporary version of the configuration files are updated. You may also make changes that will apply for future use of that machine, in which case the changes are copied back to the main version of these files on your machine.
If you have a network licence for Instat, then it is supplied for a specified number of machines. You need a registered version to be able to use Instat over a network on multiple machines.

First, install Instat on the server as for a single machine. Then, using either the CD, or the setup file over the network, to go through the installation procedure on each machine (as described below). If you are the network manager, then you may have special software to simplify this for you. Otherwise start by mapping the server folder as a network drive. You can do this in Explorer, using **Tools ⇒ Map Network Drive** Then look for where Instat has been installed on the server, perhaps in X:\Program Files\Instat for Windows\InstawS.

Now run the installation on your machine. When asked, give the folder on the network drive where Instat is already installed. Then it will see that the software is already installed and just add the local files that need to be on your machine.

You may realise that you could alternatively install the unregistered, or (free) single licence on each individual machine and avoid paying for the multiple licence. We hope that the modest price of the software will minimise this option. But, particularly for institutions in developing countries, we recognise that it takes time to ask for funds. In such cases we would not seek to delay any enthusiasts who wish to improve their statistics teaching or data analysis, while they search for funds. Sometimes initial use of the software may support such organisations to define their software strategy and hence obtain funds for other statistics packages, in addition to Instat. We consider possible strategies in Chapter 10 of this guide.
Chapter 3 - Data sets

3.1 Introduction

Many simple sets of data come in the form of ‘rectangles’. Each row of data corresponds to a case, perhaps a person that has answered a questionnaire. The columns correspond to the variables or the questions that they have answered.

Seven examples are presented that are used in many later chapters in this Guide. Ways are suggested in which the data can be organised into a form suitable for the analysis. These and other examples are supplied with Instat, so that you may try out the analyses given in this guide.

The first five examples are relatively small, and indicate the different types of data that often arise. In Sections 3.2 to 3.5 we describe a simple regression problem, a survey, an experiment, and use rainfall records as an example of monitoring data. In Section 3.6 we describe a small example where data are collected at two ‘levels’. This is a frequent problem, for example an educational study could collect data at both classroom and at pupil level.

The last two examples are larger, one is a survey and the other is an ecological study. The extra dimensions to both data organisation and analysis illustrate why many people have problems with “real” data, even if they have understood the simple examples on a training course.

Statistics packages are basically “column calculators”. Most of the dialogues are instructions to Instat, or any other statistics package, to act on the columns of data. Hence “organising” the data consists mainly of putting the data into columns. For the survey data that will be described later in this chapter, we see part of the six columns of data in Fig. 3.1a.

![Fig. 3.1a Survey data](image)

The dialogue in Fig. 3.1b uses the last of these columns and gives summary statistics for the yields that are in X6. The results are shown in Fig. 3.1c. We see more of this dialogue in Chapter 12.
The dialogue in Fig. 3.1d shows how to instruct Instat to give a histogram. This dialogue uses the same column of data and gives the graph shown in Fig. 3.1e.

We describe more about how to use a statistics package in Chapter 4. This introduction is to indicate the ease with which results can be given, once we have understood our data and know the objectives for the analysis.

3.2 A regression example - (File: regress.wor)

Fig. 3.2a shows a small set of data consisting of 10 observations (or 'rows' or 'cases') and two variables (or 'columns'). The data are from a study in Bangladesh on the cotton yield for a variety sown on 10 planting dates spaced at two week intervals from 1st September 1973. The dates of planting are coded with 1st September 1973 as day 1, and the yields are in tens of kilograms per hectare.

These data are ready for entry. The data set is small enough that the data could be entered straight into Instat as we describe in Chapter 7. The only minor decision the user may wish to make is whether the two variables should be given names to clarify the output. The data in Instat are shown in Fig. 3.2b.
As a teaching example, this is fine. Were it a real set of data for analysis we would query a study that gives such simple data. Often we find that this may be just a subset of the columns of data or that the data has been semi-processed, before being in this form. We would like studies, as a routine, to separate data entry from analysis. They should enter all the data, as collected. Then the data can be summarised into a convenient form for analysis.

### 3.3 A survey - (File: survey.wor)

**Fig. 3.3a** presents the data from a small survey on the relationship between rice yield and cultivation practice. The data are fictitious but the model used to generate them is based on the results of a regular survey conducted in Sri Lanka. The data matrix consists of the following six variables:

1. Name of village
2. Field number within the village
3. Size of field (acres)
4. Quantity of fertilizer applied (cwt/acre)
5. Variety (New Improved, Old Improved, Traditional)
6. Yield (cwt/acre)
These data are in the right “column” form for a statistics package. However most surveys are larger than we would suggest should normally be entered directly into most statistics packages. Often they would be entered into a spreadsheet, such as Excel, or special software, like EpiInfo, that is designed to support the entry of survey data. For those who use a spreadsheet, we have a “good-practice” guide titled

*Disciplined Use of Spreadsheets for Data Entry.*

The six columns of data in this survey consist of 2 that are text and 4 that are numeric. The “text” columns indicate the 4 villages in the survey and the 3 types of rice used by the farmers. So, these are not text, as in the farmer’s name. They indicate “categories” and Instat considers them as “factor” columns.
In entering factors it is often convenient to use codes. Thus the 4 villages could be coded as 1, 2, 3 and 4 and the three varieties as 1, 2 and 3. So the data may be entered as shown in Fig. 3.3b.

Fig. 3.3b Survey data as entered

<table>
<thead>
<tr>
<th>Village</th>
<th>Field</th>
<th>Size</th>
<th>Fert</th>
<th>Variety</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2.5</td>
<td>2</td>
<td>53.6</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>5</td>
<td>1.5</td>
<td>2</td>
<td>44.6</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>50.7</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>1.5</td>
<td>1</td>
<td>3</td>
<td>33.6</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
<td>5</td>
<td>2.5</td>
<td>1</td>
<td>62.1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>4</td>
<td>1.5</td>
<td>3</td>
<td>30.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4.5</td>
<td>1.5</td>
<td>3</td>
<td>25.8</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>1.5</td>
<td>2</td>
<td>1</td>
<td>61.4</td>
</tr>
</tbody>
</table>

The Village and Variety names are then entered into label columns, as shown in Fig. 3.3c so the worksheet in Instat appears as shown in Fig. 3.3d. We describe how Instat defines factor columns in Chapter 7, Section 7.3.

Fig. 3.3c Label columns  
Fig. 3.3d Data with labels attached

3.4 An experiment - (File: experi.wor)

Fig. 3.4a shows the data from a simple randomised block experiment, with 4 blocks and 3 treatments. The example is taken from Mead, Curnow and Hasted (1993) and the 3 treatments were

- O  Control (natural daylight only)
- E  Extended day (total day length 14 hours)
- F  Flash lighting (natural day + 2 x 20 second flashes per night)

The data are the number of eggs laid by a pen of 6 chickens in a 50 day period.
The layout in Fig. 3.4a is in “textbook” form and is also suitable if the data are to be analysed “by hand”. For entry into a computer package, however, the data are normally laid out as given in Fig. 3.4b. Here, one column gives the data and two further columns specify the block and treatment codes.

With experimental data, the block and treatment codes are often in a regular sequence. The sequence for blocks in this example is

1 2 3 4 1 2 3 4 1 2 3 4

and the data for treatments are

1 1 1 1 2 2 2 2 3 3 3 3

Most spreadsheet and statistics packages, including Instat, have a quick way of getting such regular sequences as these into the computer. We describe how this is done in Chapter 7, Section 7.2.

In textbooks experiments only one measurement is taken, which here is the number of eggs. In real experiments more than one measurement is taken. Often the first column gives the “unit” or plot number. Then there are the factor columns, followed by a column for each measurement.
3.5 Monthly rainfall records - (File: rain.wor)

*Fig. 3.5a* shows part of 32 years of monthly rainfall totals for Galle in Sri Lanka. This is an example of monitoring data that are collected routinely. Here the appropriate form of input may either be as in *Fig. 3.5a* or *Fig. 3.5b*, depending on the type of analysis that is required.

*Fig. 3.5a* Monthly rainfall records – Data for 1950 – 83 (56,57 missing)

*Fig. 3.5b* is appropriate if a time series analysis of the whole set of data is to be undertaken, whereas *Fig. 3.5a* is convenient if the data from each month are to be considered separately. This would, for example, be the case if a further column gives yields for a crop and one objective is to see which month’s rainfall is closely related to these yields. A
corollary from this example is that a statistics package should provide facilities to 'reshape' a set of data once it has been entered. Ways this can be done are considered in Chapter 9.

This is an example where the raw data have already been semi-processed to give the monthly totals, shown in Fig. 3.5a and 3.5b. The “raw” data are on a daily basis (or less). In such cases the analysis usually proceeds in stages, where the first stage might be to produce the monthly totals. We discuss this issue in detail in the Instat Climatic Guide, because having access to the raw data, rather than just the summary values, is often needed for a full analysis.

3.6 Data at multiple levels (File: activity.wor)

The examples above all consist of a single rectangle of data. Often information is available at multiple levels, for example a survey may be at household level, but with some questions for each person. An experiment may record yields at the plot level, but also record the height of 10 individual plants within each plot.

In such cases it is usually simplest to consider the data as two or more “tables” or rectangles, with one at each level.

A small example is shown in Fig. 3.6a and 3.6b. It represents a pilot survey, with just 8 people. At the “person level” we have recorded their sex and age group. Then they have each been asked about the activities they do. At the “activity level” we have which activity that each person does and how long he or she spent, in minutes, in the last week.

In Chapter 9, Section 9.6 we will see how to copy data between the two levels.

If we ignore the times, in x6, then there are other layouts for this type of data, which we consider in Chapter 13.9. It is called “multiple response” data, because some people have done more than one activity. This type of data is common in surveys, where it corresponds to the type of question:

“Tick which of these activities you did in the last week.”
3.7 Poverty data (Files: poverty.wor and poverty.sav)

These data were from a survey of 1789 households in Malawi and was an evaluation of part of the 2000-01 Targeted Inputs Programme (TIP) of the Government of Malawi. This project, supported by several donors, was designed to increase household food security amongst rural smallholders in Malawi. It provided rural smallholder households with one “Starter Pack”, which contained some fertiliser, maize seed and legume seed. The TIP followed on from the Starter Pack campaigns in 1998-99 and 1999-2000. They covered all rural smallholder households, providing 2.86 million packs. The 2000-01 TIP was enough for roughly half this number of beneficiaries (1.5 million).

Part of the data from the survey is shown in Fig. 3.7a. We see that the one column, X3, called RECTIP specifies whether the household received the pack.

One objective of the survey was to see whether the beneficiaries were the poorer households in the villages. Hence it was necessary to calculate a poverty index, from the basic data collected on each household, and Fig. 3.7b includes the resulting index in x30.

The need to calculate one or more indices is typical of many studies. How it was done here will be described in Chapter 8, Section 8.4. In this case, once the index was calculated, then simple tables that correspond to the main objective of the survey are of the type shown in Fig. 3.7c for the counts and Fig. 3.7d for the percentages. We describe how to produce such tables in Chapter 13. They indicate that, if the poverty index is valid, then there is no evidence that the poorer families actually received more of the support.
3.8 An ecological study (Files: coral.xls and coral.wor)

These data were from a major government study around Mauritius that began in 1992. It was to monitor changes in the ecosystem, because of concerns that the coral reef and lagoon were becoming degraded, because of pollution and other human activities.

Fig. 3.8a Transects for each site

The observations were from a series of 20 metre transects, A, B, C and D, that were at 4 random points, perpendicular to a main 100 metre transect, as indicated in Fig. 3.8a. The data were only on these sub-transects and not on the main transect. Whenever a transect intercepted some live coral, or other features, including dead coral and algae, the species was noted and the length of the intersection was also recorded.

There were 17 sets of data, and they were from 2 or 3 locations in each of 7 sites. The locations were noted as the shore-reef (SR), the back-reef (BR) that is still in the lagoon and the fore-reef (FR) that is in the open ocean. Data from one of these sets is shown in Fig. 3.8b in Excel.

Fig. 3.8b Data from one site in Excel

| Site: | Anse La Roie
| Station: |
| Transect: | A | B | C | D | Total | Total |
| Taxon | Cover | No. Cover | No. Cover | No. Cover | No. Cover | No. Cover |
| Acropora feroxosa | 375 | 2 | 15 | 1 | 2 | 20 | 1 | 690 | 11 | 1290 | 15 |
| Acropora grandis | 0 | 0 | 0 | 0 | 1 | 170 | 4 | 0 | 0 | 170 | 4 |
| Acropora sp | 0 | 0 | 35 | 1 | 0 | 0 | 0 | 0 | 35 | 1 |
| Cyphastrea chalcicira | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 |
| Fungi sp | 50 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 2 |
| Goniastrea favicularis | 0 | 0 | 0 | 0 | 70 | 2 | 0 | 0 | 70 | 2 |
| Montipora danae | 0 | 0 | 65 | 4 | 0 | 0 | 0 | 0 | 65 | 4 |
| Montipora efflorescens | 125 | 3 | 95 | 2 | 60 | 1 | 100 | 5 | 330 | 11 |
| Montipora incrustata | 0 | 0 | 60 | 2 | 0 | 0 | 0 | 0 | 60 | 2 |
| Montipora sp | 100 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 3 |
| Montipora tuberculosa | 450 | 10 | 1005 | 10 | 120 | 3 | 350 | 15 | 1335 | 46 |
| Pavona decussata | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 | 5 | 1 |
| Pocillopora damicornis | 20 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 2 |
| Seriatopora hystrix | 0 | 0 | 0 | 0 | 25 | 1 | 10 | 1 | 35 | 2 |
| Dead coral | 860 | 17 | 660 | 13 | 245 | 6 | 570 | 21 | 2335 | 57 |
| Rubble | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 60 | 1 |
| Sand | 0 | 0 | 65 | 4 | 95 | 2 | 220 | 1 | 330 | 7 |
The same data are shown in Fig. 3.8c in Instat, where the long names in Fig. 3.8b have had to be shortened.

Fig. 3.8c  Data from one site in Instat

These data typify situations where a lot of detailed data are available at a low “level”, i.e. for each 20metre transect. However the main objectives of the study will require a summary of these data so the locations and sites can be monitored.

Like the poverty study, described in Section 3.7, we will calculate various indices and describe how, in Chapter 8, Section 8.6. These include “diversity” indices, because the objectives of the study imply that the diversity of the sites should not be decreasing. We will also have to consider the “shape” of the data, because we have 17 rectangles, rather than one, and it is not immediately clear how we move from the detailed analysis at the species level, shown in Fig. 3.8b and 3.8c to the site or location level. We address these issues in Chapter 9, Sections 9.7 and 9.8.
Chapter 4 - Using a statistics package

4.1 Introduction

Instat, like almost all statistics packages, started before Windows. Then it was "easy" to use any statistics package. "All" you had to do was to "learn the language".

Many people who needed to analyse their data found learning a language to be a daunting hurdle, so they looked for alternatives. These included using a simple package that offered "statistics from a menu", or a spreadsheet, or giving the data to a statistician who knew "the language".

Those who learned "the language" for a particular package were often tied to that package, because of the investment in time they had made.

Now, with Windows, the standard approach is through menus and dialogues. As an example Fig. 4.1a shows part of the survey data, described in Chapter 3, and Fig. 4.1b shows the menu to get summary statistics.

Then Fig. 4.1c shows the Describe dialogue again, which was shown earlier in Chapter 3. There you have to put the columns you want to summarise into the data box and press the OK button. Finally Fig. 4.1d shows the resulting summaries. We see, for example that the mean yield is 40.6 bu/acre.
This “menu-based” approach is simple and we will use it through most of this guide. We describe this approach in Sections 4.2 to 4.4.

At the end of this chapter we show that the menus and dialogues are very similar in most statistics packages. So learning one statistics package is now an easy introduction to statistics packages in general and users could develop a strategy for their statistical work that involves using more than one package. We describe more about this idea of “strategy” in Chapter 10.

But, users should know that other approaches exist. The language that earlier had to be learned has not gone away. It is just in the background. For example in Fig. 4.1d the top line is

DES X6, X4

If, instead of using the dialogue, you type this line yourself, or type

: describe x6 x4

then you will get the same results.

So before Windows there was a steep learning-curve to any statistics package. You had to be able to write the language before you could start. Now you can start by just pointing and clicking, in a way that is largely familiar from other Windows software. But once you are familiar with a statistics package, then we find it is easy to progress, at least to be able to read a little of the language. In Section 4.5, we explain why you may wish to take this step.

In the final section of this chapter we describe the equivalent facilities in some other statistics packages.
4.2 Using menus

**Fig. 4.2a** Main Instat menu and File menu

<table>
<thead>
<tr>
<th>File</th>
<th>Edit</th>
<th>Submit</th>
<th>Manage</th>
<th>Graphics</th>
<th>Statistics</th>
<th>Clicmat</th>
<th>Window</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Worksheet</td>
<td>Ctrl+N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open Worksheet</td>
<td>Ctrl+O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open From Library</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close Worksheet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Import/Export</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close</td>
<td>Ctrl+S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Save</td>
<td>Ctrl+S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Save As</td>
<td>Ctrl+S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Print</td>
<td>Ctrl+P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Print Preview</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SURVEY.WOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPAL.WCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POVERTY.WOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In **Fig. 4.2a** we show the main menu in Instat, with the File menu open. Sometimes beginners to computing are daunted by the large number of different menu items, that they feel they need to understand, before they can use the package effectively.

This fear largely misses the point of using software in Windows. It is the equivalent of feeling that you need to totally master a foreign language, before you are prepared to visit that country. Even more important, it ignores the fact that you probably understand a lot of the menu items, because you have met them in other Windows packages that you use already.

As with the main menus in Instat, all Windows packages have a File and Edit menu on the left and a Window and Help menu on the right. Though their detailed contents may differ slightly, their main functions should be clear.

For example, in **Fig. 4.2a**, with the file menu, the first item is “New Worksheet” to start with a blank worksheet, in which you enter data. The second item is to open one you used previously. It would be disturbing if these options did not exist!

Some menu items have a “shortcut” associated with them. This is the key sequence on the right of the menu in **Fig. 4.2a** that allows you to open a dialog box directly from the keyboard, without using the mouse. For example, holding down the control key and pressing the letter n [Ctrl+N] opens the dialog box for a new worksheet.

Looking further down the File menu, it is perhaps comforting that there are Import and Export items, in case you start with data in some other package, or wish to continue your work elsewhere.

The other key sub-menus on the main menu are “Submit”, “Manage”, “Graphics” and “Statistics” and there are equivalents in all statistics packages. We show each in **Fig. 4.2c**, except “Submit, which is described in Section 4.5.
The Manage menu, Fig. 4.2b, is because all statistics packages need to be able to manipulate the data, make transformations and so on, so the data are ready for the analysis. Chapters 6 to 9 of this guide describe how data are entered and then organised, and this largely uses the Manage menu.

Data are often explored and presented graphically, which is the purpose of the graphics menu, shown in Fig. 4.2c. We describe Instat’s facilities for graphics in Chapter 11.

This leaves the statistics menu, in Fig. 4.2d, which is the main one in any statistics package. Most items have their own Chapter in this guide. In this menu, the first 2 items are Data Summary, covered in Chapter 12, and Tables, covered in Chapter 13. They, together with the graphics, provide most of Instat’s facilities for descriptive statistics. Many studies only require descriptive summaries, so users then do not need to venture any further in the statistics menu.

The next item is called “Simple Models”, and is described in Chapter 15. Before that we have had to introduce the ideas of “Probability Distributions”. We then move to “not-so-simple-models”, namely models with more parameters. They are described in Chapters 16 and 17. These are both quite long chapters corresponding to the importance of the next items, namely “Analysis of Variance” and “Regression” in the statistics menu. The other items are described in the later chapters.

4.3 Dialogues

Dialogues correspond to tasks, so once you know what you need to do, then it is usually easy to find the corresponding dialogue. Here we give some examples, and at the same time we show the logic of Instat’s dialogues.

We start with the Describe dialogue, for which the menu and the dialogue were shown in Section 4.1. The route to this dialogue is Statistics ⇒ Summary ⇒ Describe.
Fig. 4.3a  Statistics ⇒ Summary ⇒ Describe

In Fig. 4.3a we show again the dialogue used in Section 4.1. This is designed to be as simple as possible, and to give the “default” results we think will often be what is wanted.

If you want more, then you can ask for it, and the 3 options in Fig. 4.3a are to have more statistics, to include percentiles and to include a graph. If you tick everything, then you can get to the dialogue shown in Fig. 4.3b.

So the idea, with Instat, is that dialogues start simple, and this simple use provides minimal sensible results. If you need more, then it is easy to ask for it.

Once you press OK, then Instat remembers how you last used the dialogue. So when you return, you will usually find that the same options are still present.

As a second example, we consider the task of transforming a column of data. For example, maybe in the survey, described in Section 3.4, you wish to analyse the square root of the yields. It should be clear that this is the type of task for the “Manage” menu, shown earlier in Fig. 4.2a. Also that this is likely to be either under “Calculations” or “Transformation”.

Fig. 4.3c  Manage ⇒ Calculations

We look first at the Calculations dialogue, shown in Fig. 4.3c. Initially the OK button is disabled. This is because there is, as yet, no calculation. It will be enabled automatically, as soon as the minimum in the dialogue is completed, in this case as soon as there is anything in the calculation field. Occasionally Instat includes tooltips. They are only provided where we feel they are needed, and here it is to explain that “sqr(X6)” will give \( \sqrt{X6} \) and not \( X6^2 \).

The option to save the results is shown in the completed dialogue, in Fig. 4.3d. The alternative is to just to calculate them and look in the output window.
Although not needed for this calculation, we also have checked the “Maths” checkbox, to show other functions that could be used. Here this is mainly to show that the same logic applies to this dialogue as to those shown earlier. The default is for a simple calculator, but there are more functions available if they are needed.

All dialogues have a Help button, in case you need more guidance on how it should be used. We describe Instat’s Help facilities in Chapter 5. This dialogue is unusual in also having extra help on each of the sets of further functions.

Often there is more than one way to get the same result. With the calculator you are free to choose any formula you like, but there is another dialogue, with Manage ⇒ Transformations ⇒ One to One that could alternatively be used. This is designed for the most common transformations, and that includes taking square roots. We describe the use of that dialogue in Chapter 8.

The final dialogue we have shown is for a simple model. The route is indicated in Fig. 4.3e and the dialogue is in Fig. 4.3f. These six dialogues are used to reinforce the teaching of statistics. They are for one or two samples, and for data of different types, i.e. Normal, proportions (binomial) and Poisson. The dialogues and the display of the results are shown in Chapter 15 and are all similar. This is to encourage users to realise that ideas once learned from the analysis of one distribution, can be applied to another.

When you use some of the dialogues you may also have to consider the "layout" of your data, and we show some options in Fig. 4.3g for the dialogue shown earlier in Fig. 4.3f.

For this same dialogue, on the analysis side you also need to consider what you are estimating, with the options as shown in Fig. 4.3h.

This dialogue includes a “graph” option, which is common in the statistics dialogues. Thus graphs can also accompany a particular analysis as well as being given by graphics dialogue. In this case it produces what might be called a “teaching plot”, because it is particularly designed to support the teaching of the ideas of confidence intervals and significance tests, that are part of the results from this dialogue.
The results, other than graphs, are displayed in the “Commands and Output” window. Most statistics dialogues, as here, also have a “Save” option. This is used to save the key results back to column(s) in Instat’s worksheet. This is because the results from one dialogue are sometimes needed as the data for a later stage in the analysis.

Some dialogues have an Apply button. This is the same as OK but leaves the dialogue on the screen for further work. You can also recall the last dialogue by pressing the button on the toolbar.

Press OK, once a dialogue is completed, or use the cancel button if you change your mind.

4.4 Windows

Instat, like most statistics packages, has two main windows, one for the data and the second for the results. An example is shown in Fig. 4.4a.

Fig. 4.4a Instat’s data and results windows

If you do want to tile the Windows, then an alternative is horizontally, as shown in Fig. 4.4b.
The active Window is the top one when you tile, so if in Fig. 4.4b, you wanted the worksheet to be on the top, then make it the active window, and then tile again. Fig. 4.4b also shows the Windows menu, where you can see there are other options.

We usually find it is convenient to tile these two windows most of the time. Occasionally we have just a single window, when we wish to concentrate on entering the data, or looking at particular results. Sometimes, when other windows, described below, are also used, the tiling becomes messy. In such cases one solution is to minimise the windows you do not want to see, and then do the tiling.

As noted in Section 4.2, you can use “shortcuts” instead of menus. As shown in Fig. 4.4b, pressing the F2 key will tile visible windows vertically, while holding down the Shift Key and pressing F2 will tile windows horizontally.

If you are going to use a statistics package regularly, it is worth spending a few minutes, at some stage, to decide on the layout of the windows that you find the most convenient. You can then use the Edit ⇒ Options dialogue to make this layout the default that Instat will start with in the future.

We describe other features of the worksheet window in Chapter 6. Two features of the output (or results) window concern how to copy the results into another package and how to display the information clearly when a projector is used in a training course.
To output particular information from the results window you can mark it and copy to the clipboard in the normal way. In addition, as shown in Fig. 4.4c, there is a Copy ⇒ Special dialogue, shown in Fig. 4.4d, that lets you copy as an RTF file, or make multiple spaces into tab characters. This can be useful if you want the results to go easily into Excel, or into a table when pasted into a word processor.

In training courses when we have a projector, we use Edit ⇒ Output ⇒ Font to change the font in the results Window, into something that is clearly visible to the whole audience. Similarly Manage ⇒ Worksheet Font can be used to make the worksheet clearer.

The output window is only for numeric output. Graphs have their own special window as we describe in Chapter 11. Instat also has a special system for tables that we describe in Chapter 13. These features allow a set of graphs, or tables to be tiled separately, which we find useful, both in training courses and in data analysis. Fig. 4.4f and 4.4g show an example of a graph and a table in Instat, each in its own Window.
Finally, Instat includes its own simple editor. This is like the Notepad editor that is supplied with Windows, except that it can also handle large files. This editor may be used to look at (ASCII) data files that are to be imported into Instat, as we explain in Chapter 7. Other uses include keeping a "log-file" as a record of the analyses that have been done, as described in Section 4.5.

4.5 More than menus and dialogues

In this section we propose that most users could benefit if they knew something about Instat’s commands, or the commands of whichever other package they use. What we suggest here is both quick and easy.

In the “old days” you had to learn the some of the language before you could make full use of Instat. That took time and was particularly tricky for those who only needed to use a statistics package occasionally. It was just like learning any foreign language that you did not practice.

We are not suggesting that you do this. All we suggest is that you occasionally try a few words, or at least that you are able to understand a few words. Like all foreign languages, that is much easier and will not take much time.

Most of Instat’s dialogues generate the commands that had to be typed in the DOS versions. These commands are then sent automatically to the Instat “server”, i.e. to the old Instat. Once processed, the results are returned to the “front-end” where you see them in the output window.

First we show briefly how you could type these commands yourself. Then we consider why some knowledge of the commands might be useful, when the dialogues seem so much simpler.

We showed one way of giving a command in the Introduction to this chapter. Instead of using Statistics ⇒ Summary ⇒ Describe and completing the dialogue, you could type the command

: des x6

into the Commands and Output window as shown in Fig. 4.5a. When you press the Enter key you will get the results.
A second way is shown in the sequence from Fig. 4.5b to 4.5d. Use Edit ⇒ Edit Macro ⇒ New and a little editing window will open, as shown in Fig. 4.4c. This provides a simple editor, like Notepad, that you can use in Instat.

Fig. 4.5c  Type into a text window

Into this editor, type the command

```
   des x6
```

as shown in Fig. 4.5c. Nothing happens if you press <Enter> here, because you are just in a text editor. Now use the Submit menu, shown in Fig. 4.5d, with the option to submit the “Current Window”. The results will again be displayed in the Commands and Output window.
If you wish, as shown in **Fig. 4.5e**, you could use **Manage ⇒ Interactive** and change the Input to be interactive. Then the editing window, shown again in **Fig. 4.5f**, is titled “Interactive Editing” and when you press the <Enter> key the line of commands is executed.

Why is this information about using commands useful? Well, one possible difficulty with dialogues is getting help from other people if things go wrong. If you can tell them what commands you used, then it is much easier. A record of the commands that the dialogues have generated can be kept in a Log window and this, along with any error messages displayed in the Session window can be used to help solve your problems.

The solution can also be provided as a set of commands for your analysis. These would be in a file that you could read into Instat using **Edit ⇒ Macro File ⇒ Open**, just as you did “New” earlier. Then you run it in the same way as you ran your one-line program.

Another problem with dialogues is that all the mouse clicking can become boring, if what you are doing is repetitive. So, if you have one set of data to process, then the dialogues are fine. If you have 20 sets to process in the same way, then this repetitive mouse clicking is time consuming. It is also very odd. Computers are particularly good at repetitive tasks, so why are we doing all this repetition ourselves!

Another reason for saving commands is for auditing purposes. When doing varied data analysis tasks often and on different datasets, it is almost impossible to remember the menu clicking sequence for each task. Instead, it is much easier to open a saved text file containing the sequence of commands used for a specific data analysis.

A simple way to produce the file of commands is to use the dialogues the first time, keep a record of the commands that were generated and use them to automate the analysis for the other data sets. We describe this process and more on the use of commands in Appendix 1 of this guide.

### 4.6 Making mistakes

When you start to use Instat, or any other package, it is a good idea to make some mistakes. Some users are “naturals” at making mistakes, but if you are not, then make some deliberate errors. In **Fig. 4.6a** we have just started Instat and have deliberately changed the filename to produce one that does not exist. The message seems clear.

Having made this mistake, if you now press Cancel, there is still no open worksheet. In **Fig. 4.6b**, this is clear, partly as written in the bottom left-hand corner of the Window, but also because the Manage, Graphics and Statistics menus are not enabled.
Undaunted, we have typed “describe x6”, into the “Commands and Output” window, which we learned in Section 4.5, and which would have been fine if we had correctly opened the worksheet. Instat’s response is shown, both in the status line at the bottom of the screen, and in the main window.

If this second type of mistake is made for real, then we usually find it typifies a user who may have followed a tutorial on the software, but has not understood, or not thought enough about the logic of the way the package works. In helping such users it is important to explain the logic, rather than just to correct the mistake.

4.7 Starting with Excel

We find that many people are familiar with Excel, or another spreadsheet package. In this section we look briefly at the way that Excel is used for statistical work.

Some facilities for data analysis are automatically available with Excel, but you may have to add the “Data Analysis toolpack”. If you wish to use an add-in for statistical work, then this will also have to be installed. This installation is the equivalent of installing Instat, which we described in Chapter 2. We illustrate briefly by describing the installation procedure for our Excel add-in, which is called “SSC-Stat”.

The add-in is obtained, either from an appropriate CD, or by downloading from www.rdg.ac.uk/ssc. Copy or install the add-in files to the hard drive of your computer. Where to copy depends on the setup of your PC, but in generally copy the SSCstat files to Program files/Microsoft Office/Office/Library/Analysis. Then go into Excel and use Tools ⇒ Add-Ins. This is the same procedure you would use to add Excel’s own “Analysis toolpack”, as we indicate in Fig. 4.7a.
Use the Browse button in the dialogue shown in Fig. 4.7a, to find the directory with the SSC-Stat add-in. The file is called SSC-macros.xla. Then press OK and the SSC-Stat add-in is included in Excel, as is shown in Fig. 4.7b.

As with other installations, you only need do this step once. When you use Excel on another occasion, you will see that the add-in is automatically loaded.

In Fig. 4.7c we show part of the survey data in Excel that was described in Chapter 3 and has been used for illustration in this chapter. We see the layout of the data looks very similar in Excel, to Instat. This layout is what Excel calls “list format”. Thus the data are in columns, with the top row giving a name to each column.

One slight difference from a statistics package is that we have here used row 1 to give the names of each column. In Instat, there is a special row for the names.

To show how data can be analysed with Excel we use the SSC-Stat ⇒ Statistics ⇒ Descriptive dialogue, as shown in Fig. 4.7d. We could equally have used Excel’s toolpack from Tools ⇒ Data Analysis and chosen the “Descriptive Statistics” tool.

In the dialogue in Fig. 4.7d, we analyse the same two columns that we did earlier with Instat. We have kept the default option to put the results on a new sheet.
The results are shown in Fig. 4.7e. The way we have done this, and the display of the results, can now be compared with Instat, where we showed the same analysis in Section 4.1.

One difference is that statistics packages, such as Instat, have a different type of window for the output, compared to the data, as we explained in Section 4.4. With Excel, the results are usually on a different sheet, but it is the same type of sheet as the one used for the data. In Fig. 4.4e the sheet has been named “DescStat”. When there are many sheets it is often useful to have names, so you can easily find the one you want.

One other difference is that the results in Fig. 7.4e are dynamic. For example, if you go to the cell containing the mean yield, of 40.6, you will see that Excel has stored this value, not as the number but as the formula “AVERAGE($F$2:$F$37). If the data in the original sheet are changed, then the summaries will change automatically. In most statistics packages, including Instat, the results are not linked dynamically to the data. If the data are changed, then the new data must be re-analysed.

Many who do use a statistics package, still continue with Excel to enter and organise their data, for some simple statistics, and for graphs and (pivot) tables. These are the topics up to Chapter 13 of this introductory guide. Our main objective, in writing the SSC-Stat add-in was to encourage Excel users to combine good statistical practice, with continued use of Excel, for these areas. We have also included dialogues for simple statistical inference, in Excel, that is covered in Chapter 15 of this guide. The dialogues in SSC-Stat are sometimes similar to those in Instat, and other statistics packages, to indicate that it is then also a simple process to add a statistics package, if your data analysis needs exceed Excel’s capabilities.

4.8 In conclusion

Our main aim, in this section, is to emphasise the ease with which you can graduate to any other statistics package, once one is familiar.

In Fig. 4.8a and 4.8b we show part of the main menu, and the statistics menu for two packages called Genstat and Minitab. In Minitab, for example, you see there are menus called “Manip” and “Calc” to organise the data (like “Manage” in Instat) and then a “Stat” and a “Graph” menu.
Within the statistics menu you see, from **Fig. 4.8a**, that the first item is called “Summary Statistics” in Genstat. It is called “Basic Statistics” in Minitab, in **Fig. 4.8b**. In each case that is our route to the equivalent dialogues for presenting simple summary statistics. We use the same survey data, that has been our example in earlier sections.

Part of the dialogue for Genstat is given in **Fig. 4.8c**, and for Minitab in **Fig. 4.8d**. Here we see a difference in style between the two packages. Minitab gives very simple dialogues, while Genstat offers flexibility.

The results from this dialogue are shown for Genstat in **Fig. 4.8e** and for Minitab in **Fig. 4.8f**.
We do believe that it is easy to start using either of these packages, after gaining experience of a statistics package with Instat. What we mean by “easy” is that you would not have to spend much time learning the package itself. You could proceed quickly to the analysis of your data.

This does not just apply to Instat. If you currently already use almost any statistics package, it is now easy to combine your use of that package, with a further package, that is used by colleagues, or adds new features that you would like to exploit.

This is important, because we often see analyses where users have clearly been constrained by the facilities within a single package. It is now much easier to start by deciding on the analyses you need, and then choosing the appropriate package(s), rather than assuming that the full analysis must be done with a single package.

Many people use a spreadsheet, and particularly Excel, for their statistical work. Spreadsheets are fine for simple statistics, particularly for tabulation and graphics. For Excel, there are also various add-ins, including one from Reading, called SSC-Stat that we described in Section 4.7. Again, we feel it is easy to move from a spreadsheet to a statistics package, if the analyses you need exceed the capabilities of the spreadsheet package. One route is to use Instat, either because it suits your needs, or as a stepping-stone if you will later need a more powerful statistics package. If you wish to try a powerful statistics package straight away, you will find that many of the packages offer a demonstration version or a 30-day free licence.

If we convinced you, in Section 4.5, that learning just a little about the language would be useful, then adding the skills to read and understand the same “words” in these new languages comes almost free.
In Fig. 4.8e you see that the commands for the analyses in Genstat were:

\[
\text{DESCRIBE [SELECTION=nobs,nmv,mean.sd,sem]} \ \text{Yield}
\]

They are

\[
\text{Describe ‘Yield’}
\]

in Minitab. All very readable.

As we explained in Section 4.5, there are huge benefits from being able to read the commands. For example the poverty survey, described in Chapter 3, Section 3.7 was originally analysed using the statistics package called SPSS. Part of the analysis involved calculating indices of assets and income for each of the families in the survey, as we describe in Chapter 8. Later there was a query on how one of the indices had been calculated. This was easily resolved, because the record of the results, available on a CD, and the web, also included a “log” of the SPSS commands that had been used for the analysis. This showed the exact calculations that had been done.
Chapter 5 - Getting help

5.1 Introduction

With the advent of Windows, plus access to the Internet, there is no shortage of Help supplied with software. The problem is more to find the information you require, in a reasonable time. In Section 5.2 we explain the logic behind Instat’s Help, so you can quickly find what you need. In Section 5.3 we describe where you might find help on statistics, rather than just on Instat.

There are two main differences between the Help in Instat and in other statistics packages. The first is that we are keen to see changes in the way statistics is often taught. We describe our thoughts on teaching statistics in Section 5.4. The second is that we do not assume that you will find the solution to all your problems within Instat. So we will sometimes suggest other software, just as we do within this introductory guide.

5.2 Help on using Instat

First we have this Introductory Guide itself. There is a similar guide, if your main interest is in the analysis of climatic data. Both are available in printed form and as pdf files so you can print your own copy if you wish. Similarly there are 2 tutorials, that were described in Chapter 2 of this guide.

Using Help ⇒ Contents, shows the screen in Fig. 5.2a, and provides access to all the other parts of the Help. Each can be accessed directly from the Help menu, but Fig. 5.2a adds a short explanation of the contents.
We consider here just the two options from Fig. 5.2a titled “Getting Started with Instat” and “Dialogues, Menus and Datasets”. They complement the information in the main guides and the tutorials.

Either from the screen shown in Fig. 5.2a, or directly using Help ⇒ Getting Started brings the screen shown in Fig. 5.2b. The different options shown here provide general information, both for first-time users of Instat and for those who have used earlier versions.
The second main option, again either from the screen shown in *Fig. 5.2a*, or directly, using Help ⇒ Dialogue Guide gives the screen shown in *Fig. 5.2c*. This is now the top end of the help on each of the menus and dialogues.

Start here if you want an idea of what is in each of Instat’s menus, and then you can go further down to find out about each dialogue. Alternatively, each dialogue has a Help button, so you can start at the bottom and work upwards.

From the top level, shown in *Fig. 5.2c*, you can also see the whole list of the datasets provided with Instat, some of our teaching ideas, and so on. These are the ideas that are associated with the different dialogues, so from the individual dialogue help you will see our suggestions for that feature.
To show the "route-map", click on "Statistics" in Fig. 5.2c to give information on the statistics menu, part of which we show in Fig 5.2d. There, if we click on "Random Samples" we arrive, with Fig. 5.2e, at the same help that is available directly from the Statistics ⇒ Random Samples dialogue that we describe in Chapter 20. If we had started there, then we could click on the Statistics at the top of the help in Fig. 5.2e to go up a level. This brings us to the menu in Fig. 5.2d.

From the individual dialogue Help, such as Fig. 5.2e we can also look at the commands associated with that dialogue. That takes us into the Reference Guide. We could look for examples of the use of the dialogue, which would take us into this Introductory Guide and elsewhere. And this dialogue also has some teaching ideas associated with it.
5.3 Help on statistics

In this section we consider some places where you might find help on statistical ideas, rather than on the use of the software.

This Introductory Guide is primarily to introduce Instat, but it does discuss statistical concepts at the same time. The main resource on statistics, that is supplied with Instat, is a series of about 20 “good-practice” guides. These are short guides, mostly about 15 pages, which we hope can provide concise guidance about topics as they are needed. There are 2 overview guides, shown in Fig. 5.3a.

**Fig. 5.3a** Help ⇒ Teaching ⇒ Guidelines

The others are divided into those mainly concerned with planning (6 guides), with managing the data (6 guides), with analysis and with the presentation of the results. The titles of those on analysis and presentation are shown in Fig. 5.3b.

**Fig. 5.3b** Good practice guides concerned with analysis

Some of these guides relate to the materials described in particular chapters in this guide. For example, on analysis, the guide called “Approaches to the Analysis of Survey Data” suggests that tabulation is the main initial method, and that is the subject of Chapter 13.

The guide on Inferential statistics is particularly relevant to the fitting of simple models, described in Chapter 15. If you feel that experimental data is always analysed using “ANOVA”, which is the subject of Chapter 16, then the guide called “Modern Approaches to
the Analysis of Experimental Data” may be helpful in putting ANOVA in a more general context.

The “good-practice” guides are also available on our web site, at www.rdg.ac.uk/ssc/. There they are as pdf files for printing, and html files as well as Help files.

Instat also includes two guides on special topics. The first was written by Professor Clifford Pearce in conjunction with special software that he developed, called “GenAnova” for general Analysis of Variance. This software has now been added, with his permission, to Instat. The guide is called “The Statistical background to ANOVA” and is an excellent introduction to the general concepts. It is aimed for non-statisticians, but will also give some useful insights to statisticians, particularly if they have to teach this subject.

This is accessible as one of the good-practice guides, or from the Help associated with the Statistics ⇒ Analysis of Variance ⇒ General dialogue. The use of this dialogue is described in Chapter 16, Sections 16.3 and 16.10 and also Chapter 21, Section 21.4.

The second special guide is on log-linear models and was written by Joy Merrett, who also wrote that section of Instat. It is partly intended for those who still use mainly chi-square tests to process qualitative data that is in the form of contingency tables. It is a useful guide in its own right and also because it symbolises the way statistical methods have advanced, compared with analyses that can be done “by hand”. It is available from the Help associated with the dialogue, namely Statistics ⇒ Regression ⇒ Log-Linear. The use of this dialogue is described in Chapter 18, particularly Section 18.5.

5.4 Help on teaching or learning statistics

We believe that sensible use of the computer can help to make statistics courses more accessible for users and also more fun to teach. But so far, statisticians have not been very successful, particularly in training courses to non-statisticians. Most recipients still dislike and often fear “statistics” just as they used to, in pre-computer days.

We believe that this is an appropriate time to change. In the early days we had to teach people how to use a calculator for statistical work, then we had to teach them the language of a statistics package. Now that it is so easy to use statistical software, we can concentrate on teaching statistical ideas.

What we would like to see is imaginative changes in the syllabus, so users are better prepared for their real-life needs in statistics. This does not mean a more complicated course. It is usually a broader course. There is more time for planning and design, some time for instruction on managing data and so on. Under Teaching ⇒ Strategy we give some of our ideas, and show part of that screen in Fig. 5.4a.
Fig. 5.4a  Help on teaching statistics

The Teaching of Statistics

In promoting the use of the computer to support the teaching of statistics, we are not suggesting that all statistics teaching should relate to the computer, nor that practice of hand calculations should stop. In this section we explore the following topics:

Changes in what is taught

Making changes, but not with Instat-

Using other software in addition to a statistics package

Case studies

Good statistical practice

We have also provided teaching ideas for users as they browse through the Instat+ dialogue help.

Fig. 5.4a mentions the good-practice guides again, that were outlined in the last section, and which we use within some of our training courses. There is also a set of 10 small case studies, each describing a way that access to the computers can change our teaching of statistics.

In Fig. 5.4a we also have a section titled “Using other software in addition to a statistics package”. An example, that may become available from within Instat, and is certainly on our web site, www.rdg.ac.uk/ssc is a set of “statistical games”. We show some screen shots, from one of these games, which simulates the conduct of a multi-stage survey.

Fig. 5.4b  Introduction to the rice survey game

In this survey, users sample villages, then fields within the villages and then plots within the fields. The villages are of different sizes as indicated in Fig. 5.4c, measured either by their area, or the number of farmers, and this information is available to allow a variety of sampling schemes.
Once the data are collected, they give rise to the same type of data shown earlier, in Section 3.3 of this guide. We will use that example in further chapters to illustrate aspects of analysis.

Two points are sometimes raised concerning the use of these “simulation games”. The first is that students may enjoy this game, but they learn about the computer, rather than about the design of a survey. Our answer is to have a “hand-version” of this game also, where each of the 160 fields is an envelope. The statistical challenges and the data are identical, but students sample from the envelopes rather than from the screen. We continue to use both versions, sometimes in the same practical class.

The second point is whether it would not be better for students to do a real survey, rather than a simulation? We would claim that these are not mutually exclusive. We also do real sampling exercises. This simulation can be done easily within a 2-hour practical class. This seems an effective way to teach many sampling ideas. And that applies whether students also do a real survey, or not.

### 5.5 In conclusion

Further information on all the topics introduced in this Chapter is available in statistics books, on the web, and in the Help provided by other statistics packages. We consider some of these sources briefly.


On the web, Statsoft, [www.statsoft.com](http://www.statsoft.com) the producers of the statistics package, STATISTICA, maintain a comprehensive textbook, that can be downloaded or viewed online. They also include a glossary of statistical terms.

CAST (Computer-Assisted Statistics Teaching) is an interactive course in introductory statistics. Students can choose between general, biometric or business for lectures with large graphics that are suitable for projection. CAST makes extensive use of dynamic graphs to support the explanation of concepts.

The web is a rich source for materials and one place to start for statistics teaching is at [www.ltsn.gla.ac.uk](http://www.ltsn.gla.ac.uk). This is the host of the statistics page for the “Learning and Teaching
Support Network”. This is particularly for further education, while the Royal Statistical Society Centre for Statistical Education, on [http://science.ntu.ac.uk/rsscse/](http://science.ntu.ac.uk/rsscse/) has a broader remit. In the USA, the American Statistical Association’s page, on [http://www.amstat.org/education/](http://www.amstat.org/education/) provides excellent information plus further links, while [http://www.amstat.org/publications/jse/](http://www.amstat.org/publications/jse/) gives information on the Journal of Statistical Education. Our own web site on [www.rdg.ac.uk./ssc/](http://www.rdg.ac.uk./ssc/) also provides some resources and links to other sites.

In Chapter 4 we showed that the simple use of many statistics packages was similar. This is not the case for the Help that each package provides. Hence, if you are undecided between different commercial statistics packages, then the value and ease of use of their Help may assist in your choice.

For example, in Section 4.8 we showed a Minitab dialogue to produce simple descriptive statistics. The output included the “TrMean”, short for the “trimmed mean”. In Fig. 5.5a we show the Help in Minitab that explains what is meant by the trimmed mean.

**Fig. 5.5a Help in Minitab showing the explanation of statistical terms**

![Minitab Help Screen](image)

Many packages include an on-line tutorial to introduce the basic use of the package, with SPSS providing a good example, see Fig. 5.5b and 5.5c.
Fig. 5.5b  Introduction to SPSS Tutorial

Welcome to the SPSS Tutorial. This tutorial will show you how to use many of the features available in SPSS. It is designed to provide a step-by-step "hands-on" guide. All the files shown in the examples are installed with the tutorial so you can follow along, performing the same analyses and obtaining the same results shown here. (You don’t have to -- but you can if want to.)

It is also designed to make it easy for you to start and stop anywhere you want. If you only want to learn about a few specific tasks, use the Contents and Index buttons (next to the Back and Next buttons) at the bottom of this window to find the information you need, or simply click the Next button (arrow pointing right) to step through all the tutorials from beginning to end.

Fig. 5.5c  Contents of SPSS tutorial

In addition, many packages are including their full instruction guides as files that can be examined on-line.
Part 2 – Organising the data

This second part consists of Chapters 6 to 10. In Chapters 6 and 7 we describe alternative ways of setting up an Instat workshop and entering the data.

These include typing the data directly, or importing from elsewhere. We also have a “library” of datasets, including all the examples used in this guide.

In Chapters 8 and 9 we describe common ways that data are often manipulated before the analysis. Chapter 8 describes transformation, like calculating indices, that add columns to the original data, but do not change its shape. We often do need to change the “shape” and that is covered in Chapter 9.

In Chapter 10 we give our views on alternative “strategies” for statistical work. We consider strategies for statistical software, for organising data, for analysis and for learning statistics.
Chapter 6 - Worksheets

6.1 Introduction

In Instat the data are stored in a special kind of file called a worksheet. We explain here how to create a new worksheet and how to use an existing worksheet. We also show how to get summary information about the structure of a worksheet.

6.2 Creating a new worksheet

The File ⇒ New Worksheet menu and dialogue, shown in Fig. 6.2a and 6.2b, is used to create a new Instat worksheet file.

You may then specify the maximum dimensions of the worksheet. The default is for it to contain 20 columns, each of up to 50 rows, as shown in Fig. 6.2b. These values may be altered for any given worksheet, as shown in Fig. 6.2b, but there are limits.

The most important part of the worksheet is the data matrix and the current limit is 127 columns. The length cannot be more than about 11,000 rows.

If you consistently want worksheets of a different default dimension to those provided, then use Edit ⇒ Options and change the values there. If you want to change the dimensions of an existing worksheet, the Manage ⇒ Resize Worksheet dialogue may be used.
6.3 Using an existing Instat worksheet

The **File ⇒ Open Worksheet** dialogue gives the standard open file dialogue that will be familiar from other Windows software. We show an example, in **Fig. 6.3a**.

*Fig. 6.3a*  **File ⇒ Worksheet dialogue**

This dialogue is also used to import data in a range of other formulas, as we describe in Chapter 7.

When opening an existing worksheet a backup copy is made if set in the **Edit ⇒ Options** dialogue. This is often advisable, because Instat automatically updates the working version on the disc as changes are made. There is currently no “undo” on such changes, so having a backup is a useful safeguard.

There is also a **File ⇒ Open from Library** dialogue, as shown in **Fig. 6.3b**. This looks the same, but starts in a different directory.

*Fig. 6.3b*  **File ⇒ Open File Library dialogue**
The worksheet library contains sample worksheets, including all those used in this guide. If you choose that option, then Instat will make a copy of that worksheet in your current directory, and then open this copy. On future occasions you continue to use the copy in your current directory. You would only use the library again, for the same worksheet, if you have messed up your own copy.

### 6.4 Using Instat’s worksheet editor

Instat has a special worksheet editor, which is accessed using **Edit ⇒ Run Worksheet Editor**, as shown in **Fig. 6.4a**. An example of the editor in operation is shown in **Fig. 6.4b**. Within the editor you may create new worksheets, just as in Section 6.2, or open existing ones as in Section 6.3. You can also enter data into these worksheets and copy information between them.

**Fig. 6.4a  Edit menu**

Within the worksheet editor the files are not updated on the disc when changes are made. This editor behaves as a standard Windows program so use **File ⇒ Save** or **File ⇒ Save As** when you wish to save information back to the disc.

In **Fig. 6.4b** notice that the worksheet editor has just the File, Edit, Windows and Help menus. It is just an editor and is not used to analyse the data. You should also not have the same file open in Instat and the editor at the same time.

### 6.5 The data in a worksheet

To provide an example we create a new worksheet, as shown in Section 6.2. We call the worksheet Temp, for “temporary”, because we only want it for illustration. Then we use the **Statistics ⇒ Random Samples** dialogue, as shown in **Fig. 6.5a**, to enter some data into the first 5 columns. We describe the use of this dialogue in more detail in Chapters 20 and 21, but here we just want a quick way to enter some numbers.
Now we use the Manage ⇒ View Data ⇒ Format dialogue as shown in Fig. 6.5c. There we have declared that the first 10 columns, x1-x10 should all be displayed to one decimal place only. The result is shown in Fig. 6.5d.

You might have found it odd that we gave a format to x1-x10, when we only had data in x1-x5. This was just to show that columns could be “prepared” for data that will be supplied later. If you now use the Statistics ⇒ Random Samples dialogue again, as shown in Fig. 6.5a, but using x6-10, then the data will be displayed automatically with one decimal place.

The changed look to the data between Fig. 6.5b and 6.5d affects only the display and not the data themselves. They are still stored in Instat with full precision.

Statistics packages are effectively “column calculators”, so the data columns x1, x2, … are the most important part of a worksheet. Sometimes it is useful to store other data structures also, and these are now described. You can see in Fig. 6.5b and 6.5d, that the worksheet
has four tabs, labelled Columns, Constants, Strings and Labels and these are considered in turn.

In Fig. 6.5e we show the use of the Manage $\Rightarrow$ Calculate dialogue that is described in more detail in Chapter 8. Here we show an example where the mean of the data in column x1 is calculated. The mean is a single number and we choose to store it in a structure called K2. We show the result in Fig. 6.5f, where we have chosen the Constants tab. We had already entered a number into k1. We see in Fig. 6.5e, under Data, that we can use what is stored in k1, k2, etc. in further calculations, so it just saves us from typing these individual numbers each time.

In Fig. 6.5g and 6.5h we show the two other types of data that can be stored in a worksheet. Strings can be used to store simple formulae and we will see, in Chapter 11, how these functions can be plotted. They can also store descriptive text, like the title of a graph.

Label columns, as shown in Fig. 6.5h are used to store columns of labels that will be used to label tables of means or percentages. They partly make up for the fact that Instat’s main data columns can only store numbers. When the data column is of categorical data, perhaps recorded as 1, 2 or 3, then these labels can be attached so the data are clearly seen to represent different categories. These are called factor columns in Instat and we show how they are set up in the next Chapter.
6.6 Navigating round the worksheet

The features we describe in this section apply equally to the active worksheet in Instat and to any worksheets in the Worksheet editor, that we introduced in Section 6.4.

The keys and use of the mouse to move around within a worksheet is similar to any spreadsheet package. We use the survey data, shown in Fig. 6.6a, as an example.

If you use the mouse to move about, then click or use the sliders to show different parts of the sheet. To mark any section, start in the cell that is in one corner of the part you want to mark, and drag the mouse to the other corner. Alternatively click the left button in one corner, then go to the other corner and then do <Shift> <Left-Click>

With the keyboard, the arrow keys move the current position, one cell at a time. Pressing <Ctrl> while you use an arrow, moves you to the beginning, or end of the columns or rows. If you have the <Shift> key pressed, while you use the arrow keys, then it marks the block as you move.

The <Home> key moves you to the first column, and <Ctrl> <Home> moves you to the top left hand corner of the sheet. Similarly <End> and <Ctrl><End> move you to the last column and to the end of the data in the sheet.

Sometimes it is useful if the first few columns are always visible. If the cursor is at the start of the data, a small padlock shape appears. This may be moved and in Fig. 6.6b we show the first column that we have frozen in this way. Now when you scroll sideways, the first column will always remain visible.

The width of columns and the height of rows can also be altered, and we have used this feature frequently in this guide particularly to show more columns of data in some of the figures. The Manage ⇒ Worksheet Font dialogue may be used to change the font, size and colour of the complete worksheet. This is useful for presentations when the default may be too small. This has been used in Fig. 6.6b.

We describe entry of data into the worksheet in Section 7.2, and the editing of data in Section 7.8.
6.7 The information about a worksheet

Sometimes we need to know about the data in the worksheet, rather than needing to view the actual numbers. We use the **Manage ⇒ Worksheet Information** dialogue for this, as shown in **Fig. 6.7a**. The information, for the survey worksheet, is shown in **Fig. 6.7b**.

![Fig. 6.7a Manage menu](image)

![Fig. 6.7b Manage ⇒ Worksheet Information](image)

In **Fig. 6.7b** we are looking at the summary information about the different components of the worksheet. For example we see that there are 20 columns available, of which we have used 7. If we are going to need more than the 20 columns in our work, then we could use **Manage ⇒ Resize Worksheet**, to change the dimensions.

We can also use the options at the top of **Fig. 6.7b** to look at individual components of the Instat worksheet. For this worksheet we look at the information on the columns in **Fig. 6.7c**. For example, we see that we have data in x1 to x7, and that they are of length 36.

![Fig. 6.7c Information about columns](image)
6.8 In conclusion

You may find it odd that you need to give Instat the name of the file, when the new
worksheet is created, rather than when you first save it. The reason is that Instat
automatically updates the contents of the worksheet on the disc, whenever its contents
change. When Instat was first developed, this was a major strength, because the worksheet
was always up-to-date, even if there were occasional power cuts.

Now this feature is a mixed blessing. It is still an advantage that the worksheet remains up-
to-date, and has helped many users, when they have had problems, while they were using
Instat, and had perhaps to reboot the machine. On the down-side is the fact that this is not
current Windows philosophy. More important it means that there is no “undo” operation on
the contents of a worksheet. Hence, as explained in Section 6.3, once you open a
worksheet, there is an option to declare a backup as a safeguard

Worksheets in Instat are limited, by comparison to spreadsheets and more powerful
statistics packages. Some spreadsheets have a limit of 256 columns, compared with the
Instat limit of 127. The number of rows may be limited to 65,000, compared to about 11,000
in Instat. Many of the statistics packages do not have a limit. If you have a huge dataset
then you need a large hard disc, and perhaps patience, but that is all.

Spreadsheets and standard statistics packages are also more flexible in the data that can be
stored in the columns. Usually there can be character or text columns, or date columns as
well as strictly numeric columns.
Chapter 7 - Entering the data

7.1 Introduction

In Chapter 6 we showed how to create an Instat worksheet. Now we explain how data can be entered. They may be entered directly, as described in Section 7.2, or imported from elsewhere, as we describe in Sections 7.4 to 7.6.

In Section 7.3 we show how to define “category” or “factor” columns, and also how to attach labels to each category, or level of the factor. In the remaining sections we first show how to lock the columns, so the data are not accidentally overwritten. We also describe how to transfer data between Instat worksheets and how to export data from Instat.

7.2 Putting data into the worksheet

Once an Instat worksheet has been created, as shown in Section 6.2, the simplest way of data entry is just to type the data as into any other spreadsheet display. Then optionally we add the names at the top of the columns.

The regression data, introduced in Chapter 3, Section 3.2, are shown again in Fig. 7.2a and they could be entered in this way. When typing the data, you can choose to enter by column, or by row. If you go “down”, then press the <Enter> usually <↵> on the keyboard to move down to the next row. Pressing <Ctrl><↵> takes you to the top of the next column. If you go “across”, then use <tab> usually <⇥> on the keyboard to go to the next column.

In Fig. 7.2a, X1, which is an example of a regular sequence, could be entered more simply. The numbers increase by 15 and Instat has a special option for this. It is Manage ⇒ Data ⇒ Regular Sequence, and we show the dialogue in Fig. 7.2b. The dialogue includes a “Preview” feature, shown in Fig. 7.2b, so you can check that what you generate is what you want, before pressing OK.

Fig. 7.2a Data

Fig. 7.2b Manage ⇒ Data ⇒ Regular Sequence

Instat’s worksheet editor was described in Section 6.4. This is accessed using Edit ⇒ Run Worksheet Editor and you may type data there instead. If you use the editor you may type or edit data just as described above. But you do not currently have the facilities to enter regular sequences, nor to define factors.

The editor allows multiple worksheets to be open at the same time, so it is easy to copy information between them, or between one worksheet in the editor and the open worksheet in Instat.
An Instat worksheet includes four different types of data, as described in the previous chapter, Section 6.5. These are data columns (X1, X2…), label columns (L1, L2…), Constants, (K1, K2…) and Strings (S1, S2…). Each is entered in the same way. A title can also be given to the worksheet.

### 7.3 Defining factor columns

**Fig. 7.3a** shows part of the survey data, described in Chapter 3, once they have just been entered. In the current version of Instat, the columns, shown in **Fig. 7.3a**, can only be numeric. Here, both village, X1, and variety, X5, are “category”, or “factor” columns. There are 4 villages in the survey and 3 varieties. We can enter their names, up to 8 characters, into label columns in the worksheet, as shown in **Fig. 7.3b**.

**Fig. 7.3a** Data before defining factors

**Fig. 7.3b** Label columns

![Data and Label Columns](image)

Now we use the **Manage ⇒ Data ⇒ Factor** dialogue as shown in **Fig. 7.3c**, to define the data columns X1, and X5 as factors and attach the labels from the label columns, L1 and L2. For the data in column X1, this labels all 1’s with the label in row 1 of L1, i.e. SABEY, and so on. Once this is done the display of the worksheet changes automatically as is shown in **Fig. 7.3d**. The column names in **Fig. 7.3d** have an F added, as in “X1 – F” and the labels are automatically attached.
The simple experimental data, described in Chapter 3 and shown again in Fig. 7.3e, combines both the features above. The factor columns are in a regular sequence, as illustrated in Fig. 7.2b. And the treatment column needs to be labelled, as was illustrated for the survey data in Fig. 7.3b and 7.3c.

7.4 Importing data

Often the data will already have been entered into another package. This may be into a spreadsheet, such as Excel, or into another statistics package, such as Minitab, or Genstat.

The example in Fig. 7.3a was prepared, in Microsoft Excel, for a training course that used the statistics package called Genstat. Features to note are that the data start in the cell A9, and that some column names finish with a "!". This is so they import into Genstat directly as factor columns and the same applies to Instat.
In Fig. 7.4b we show Instat's File menu and here we use the File ⇒ Open Worksheet dialogue again. The difference from Chapter 6, where we opened an existing Instat worksheet is mainly that we choose the appropriate file type, as shown in Fig. 7.4c. Many different formats can be imported into Instat, and just a few are shown in Fig. 7.4c. The list is not exhaustive and the importing works by recognising the file extension.

In looking at the menu in Fig. 7.4b, you will see there is a special File ⇒ Import/Export submenu and you may be surprised that this is not used for the importing here. It is described in the next sections when we consider the importing of data from ASCII files, and a facility called ODBC, but the simple importing now uses the standard File ⇒ Open Worksheet dialogue.
Pressing the Open button on the dialogue in Fig. 7.4c gives an import dialogue as shown in Fig. 7.4d. There the default name of the Instat worksheet that will be created is the same as the Excel name, but with the Instat extension. In Fig. 7.4d we have also specified that we only wish to import the data part of an Excel sheet, by accepting the “Named range”.

In this case the data are imported into a new Instat worksheet that is specially created. The import has included the column names and also defined the factor columns, as shown in Fig. 7.4e.
7.5 Importing an ASCII file

An ASCII file is one that can be read into a text editor, such as Notepad. In pre-Windows days one often transferred data by exporting, as an ASCII file, from the first package and then reading this file into the second one.

We show various ways that a simple ASCII file can be imported into Instat. An alternative is to import into something else, like Excel, that has good facilities for importing ASCII files. Then the data can be transferred to Instat using the method described in the previous section.

With an ASCII file it is usually useful to look at the file first, just to check on the layout of the data. We illustrate with a simple example from Mead, Curnow and Hasted (1993), page 10. Instat has its own text editor and so the data can be examined using Edit ⇒ Edit Text ⇒ Open. Then browse to look for the file, which is shown in Fig. 7.5a. We see there are 10 columns of data, though the description informs us that these are 100 observations from a single variable. If the file is in the Instat library, then it is useful to save a copy in the current working directory.

Fig. 7.5a  ASCII file to import

The first step in importing the ASCII file into Instat is to use File ⇒ New Worksheet as indicated in Fig 7.5b, to create an empty worksheet that is ready to receive the file. In this case the data are eventually to be in a single column, so the worksheet must have column length at least 100.

A simple way of transferring the data is to use Copy/Paste. In the editor, just block the data and use Edit ⇒ Copy, or <Cntl>C. Then go to the first cell in X1 in the Instat worksheet and use Edit ⇒ Paste, or <Cntl>V. Even better would be to paste the data starting from column X2. This is because we want the data to be in one column of length 100, and not 10 columns of length 10. This “stacking” of the data is easily done, using Manage ⇒ Transformations ⇒ Stack, which we describe in Chapter 9.

There is also a special File ⇒ Import/Export ⇒ Import ASCII dialogue, and there are 3 reasons why we might want to use it, rather than the Copy/Paste. The first reason is that the above layout of the ASCII file was simple, because there were spaces between each number, so it transferred easily into separate columns. This is an example of “fixed-format” data, because each number, like 13.2, has 4 digits. Sometimes there may not be spaces, for example, the top line of data may be given as

13.212.616.214.313.417.818.815.618.0 7.6
and then Copy/Paste can not be used. Notice in the above line there is a space before the 7.6, because each number has 4 digits.

The second reason is that we may have many files to import, that all have the same structure. Then it is better to experiment with the first file, until the importing is exactly as you would like it. Then record a “macro” so the other files can be imported easily. We mention writing macros in Appendix 1, but it is described in more detail in the Climatic Guide, Chapter 14, and is also used there, in Appendix 3, where the importing of a variety of awkward data sets into Instat is described.

The third reason is that we might want only to import part of the data in the ASCII file.

The first illustration of the File ⇒ Import/Export ⇒ Import ASCII dialogue is shown in Fig. 7.5c. Here we have used the option to import into a single column and this is an advantage over the Copy/Paste method. The result is shown in Fig. 7.5d.

For the second and third examples, we assume that the 10 columns in the ASCII file in Fig. 7.5a are as we require, but that we only want to import selected columns. The example in Fig. 7.5e, shows how to import just columns 1, 3, 5, 7 and 9 into X2 to X6.
Finally we show how to import fixed format data. Each “field” in the ASCII file shown in Fig. 7.5a has 6 characters, 2 of which are spaces. This information is given to Instat in Fig. 7.5g and here we have chosen to just import the first 5 columns.

This third example would have worked, even if there were no spaces between the numbers. Using this third method, the dialogue has also generated a number of Instat commands that are in the Commands/Output window and are shown in Fig. 7.5h.

These commands are the beginnings of a macro that could be developed to import data in this format. Those who are experienced in computing may wish to follow this idea of writing macros further. Beginners should be aware that the idea of writing macros for such tasks can save time on repetitive tasks.
7.6 Using an ODBC query

ODBC stands for Open Data Base Connectivity. This is a more flexible way of importing from external data sources. It can include features typical of database queries, like importing a subset of the rows of a large dataset. To import a subset we give criteria that the data must satisfy.

We give an example for an Excel workbook, but this facility can also be used to import data from dBase, FoxPro, Access files and SQL servers. We use the same Excel file shown on Fig. 7.4a but assume that the workbook has multiple worksheets and named ranges.

To show the flexibility of this system we import just four columns of data and just those rows that do not have the 'Control' treatment and which also have values of the variable 'quantity' in the range of 100 to 110 units.

Use the File ➔ Export/Import ➔ ODBC Query dialogue (Fig. 7.6a) and click on the tab labelled “Machine Data Source”.

This tab will show which types of data sources you have currently installed on your PC. You can add more types by pressing the “New” button and adding a new Microsoft ODBC driver for example, Oracle data sources. Select Excel and click OK.

Then, in the Select workbook dialogue that follows, Fig. 7.6b, locate the folder that contains the Excel workbook in question and double-click on the file name.
In the Open Spreadsheet dialogue (**Fig. 7.6c**), the pull-down list indicates all the tables of data in the workbook. In this list, the caption SYSTEM preceding a name indicates it is a worksheet, whereas the caption TABLE, shown in **Fig. 7.6c**, indicates it is a rectangular set of data. They can either be in Excel’s “list format” or be a named range.

When a table is selected, a list of its columns is displayed in the “Available Columns” box. Column names are prefixed with a letter giving their type: C for text, N for Numeric and D for Date. Date values are read as text, as ODBC databases do not have a standard method of storing date values.

Select the TABLE: Data and notice that the column names ending with a ! in the Excel sheet are now followed by an underscore, so the explicit meaning of this character is lost. Thus column treat_ will be imported as a factor because it contains text, but the column called farmer_ will be imported as numeric (see **Fig. 7.6f** and compare with Section 7.4) and will have to be declared as a factor later (Section 7.3).

Select the four columns as shown in **Fig. 7.6c**.

Selection of columns from a table creates a SQL SELECT statement, which can be expanded if you know how to write SQL statements. In that case tick the Edit resulting SQL statement checkbox.

Next, we select a subset of the rows. If you wish to do this, click the “Where” button in **Fig. 7.3c**. This opens the dialogue shown in **Fig. 7.4d**, which can be completed as shown. Notice that factor levels are included in single quotes.
Click OK to return to Fig. 7.6c and Ok again to run the query. You should get a message as shown in Fig. 7.6e, and then be asked for the name of a worksheet, where you wish to save the data.

Now use File ⇒ Open Worksheet and open the worksheet that has been produced. The result is shown in Fig. 7.6f.

At the bottom of the desktop the ODBC tab indicates this application is still active. ODBC is a component added to Instat, so every time you run a new query through the Instat menu a new tab appears at the bottom of your desktop. Instead of opening several ODBC tabs, you run a new query by a left click on the current tab and select Query ⇒ Database.

You can also save the SQL query by ticking the “Save retrieval in file for re-use” checkbox in the Open spreadsheet dialogue (Fig. 7.6c). You will be prompted for the name of an Instat Stored ODBC retrieval (.IDB) file after your query has run correctly. The current ODBC database specification and SQL statement will be stored in this file, and may be rerun at a later date by selecting Query ⇒ Rerun saved in the Instat ODBC query dialogue. The saved *.idb file contains statements such as those shown in Fig. 7.6f.
7.7 Protecting data

Some security is automatically provided for the data in your worksheet, unless you have used the WARNings command to turn warnings off. Any attempt to overwrite data will prompt Instat to ask whether you really want to lose the original data. To give extra protection use the Manage ⇒ Data ⇒ Lock/Unlock dialogue as shown in Fig. 7.7a.

Once locked, a column cannot be overwritten or edited in any way. The protection can be lifted by the unlock option in the dialogue.

In the display from the Manage ⇒ Worksheet Information dialogue, described in Chapter 6, the column headed ‘State’ tells you whether columns are unlocked or locked.

Sometimes Instat locks columns automatically, without being told to do so. The columns are then said to be ‘system-locked’. This happens when one data structure is derived from others, or attached to them in some way. If the original columns were changed, then it might be possible to get contradictory results. For example, in Fig. 7.3c and 7.3d, the label columns L1 and L2 were used to declare factors. If the labels were altered, then the definition of the factors may no longer be compatible with the labels. Thus, L1 and L2 have been locked by the system. This is shown in Fig. 7.7b, where the labels have a “*” after them.

The only way to lift the protection imposed by system-locking is to remove or undo the actions which caused the locking in the first place. For example, in Fig. 7.7b, if we were to remove the factor X1, or change it from being a factor, then L1 would no longer be system-locked.
7.8 Editing data within a worksheet

The simplest way to edit data in a worksheet is usually to type the correct values in the required cells. Left-click, in the cell, to retype. Double-click to edit the current contents of the cell.

To move data around, you can also use the Editing options, Cut/Copy/Paste, as in any Windows program. This is also currently the simplest way of inserting a blank row or column in a worksheet.

Editing data in a factor column can be done in two ways. You can use the pull-down list, that gives all the options for that factor. Or you can reset the column back to being an ordinary variable, using Manage ⇒ Data ⇒ Factor as shown in Fig. 7.8a. Then the column becomes numeric again, and you can type the correct numbers. Then use the Manage ⇒ Data ⇒ Factor dialogue again, once the editing is complete.

If you want to edit the labels that are currently used by a factor column, then you will find that Instat has locked them, for the reasons explained in the previous section. The solution is to use the Manage ⇒ Data ⇒ Factor dialogue first, as shown in Fig. 7.8a, and reset the factors back to ordinary variables. This automatically unlocks the label columns and they can now be edited.

Use Manage ⇒ Duplicate (Copy Columns), as shown in Fig. 7.8b, to copy whole columns and retain all the information about factors.

As Instat worksheets have limited space, it is usually useful to delete unwanted data columns. This can be done using Edit ⇒ Cut, but there are two alternatives. The first is to use the Manage ⇒ Remove (Clear) dialogue as shown in Fig. 7.8c.

This dialogue is usually used to remove columns of data, but it can equally be used to clear other types of data too. It can also be used to remove rows of data. With the rows there are 2 options, which are either to move all the remaining rows up, or to insert missing values.
The second alternative is to type the corresponding command into the Commands/Output window. In the above example this is just

: remove x8-x10

or even shorter as

: rem x8-x10

and is often much quicker than using the dialogues.

### 7.9 Transferring data between Instat worksheets

A simple way of copying a whole worksheet is, as with other Windows software, to open the worksheet to be copied. Then, with the worksheet the active window, use File ⇒ Save As. Give it a different name. Instat then makes a copy of the current worksheet and automatically opens the new worksheet it has just made.

Data can be copied between worksheets via the clipboard, as is usual with Windows software. Instat can only have one active worksheet at a time, but there are two ways round this problem. The first is to copy from the first sheet to the clipboard, using Edit ⇒ Copy. Opening the second worksheet will automatically cause the first one to close. Then use Edit ⇒ Paste.

Alternatively use Edit ⇒ Run Worksheet Editor. Within the editor you can either open one of the worksheets and copy to the one in Instat. Or the whole operation can be done in the worksheet editor, because the editor has no restriction on the number or worksheets that are open at the same time.

In Fig. 7.8b we saw the use of the Manage ⇒ Duplicate (Copy Columns) dialogue to copy columns within a worksheet. This same dialogue may also be used to copy columns from another worksheet.
7.10 Missing values in Instat

Each worksheet may have its own missing value codes. Usually there is just a single code, but up to three are allowed. The default is for the value –9999 to be used as the missing value code, but you can change that for new subsequent worksheets with the Edit ⇒ Options dialogue. The Manage ⇒ Resize Worksheet dialogue is used if you wish to change the codes missing value in the current worksheet.

The spreadsheet denotes a missing value by * (or **, and ***, if you use more than one different code). If you are entering data, then a missing value may be typed as a *, or the value of the code, e.g. –9999 may be typed. When data are imported from other applications, missing values, like blank cells in Excel, will automatically be recoded to the missing value and denoted by a *.

Most of Instat's dialogues and commands cope with missing values in the obvious way, namely by leaving out the corresponding observation. There are some special situations and we describe how Instat copes with missing values in tables, ANOVA and regression, in Chapters 13, 16 and 17.

Sometimes you may need to refer to a missing value in a calculation, or to select a subset of the data. Then you cannot use * as missing, because Instat confuses it with * as multiplication. Instead you can use M(0) as the missing value (with M(1), and M(2), if you have more than one value).

7.11 Exporting Instat worksheets to other applications

You might want to export Instat worksheets to other applications that have more powerful statistical modelling techniques graphs. We illustrate by exporting the survey worksheet to an Excel workbook.

Use the File ⇒ Import / Export ⇒ Export as, dialogue

In the worksheet from which the data will be exported box, this box is filled automatically, see Fig. 7.11a if you wish to export the current worksheet. Use the bottom Browse button to specify the folder where you want to export the worksheet. This also lets you specify the File type from a pull down list: select Excel, specify the filename and click OK.

Fig. 7.11a  File ⇒ Import/Export ⇒ Export

- Worksheet: C:\MY\DCCU\WORKING\SLAYEY
- File to which data will be exported
- Export file: C:\My Documents\working\survey.xls
- Help
- Reset
- Cancel
- OK

Sometimes you may wish to export the Instat data as an ASCII file. This can be done, via the dialogue above. An alternative is to use the File ⇒ Import/Export ⇒ Output, dialogue. This only permits the output of ASCII files, but it does have the flexibility of permitting the output of selected columns, and can also add the necessary commands so the resulting file can easily be read into some of the common statistics packages.
7.12 In conclusion

Instat, like most statistics packages, has no pre-programmed facilities for data checking. Hence, unless the data set is small, it is usually desirable to enter it elsewhere, and then import into the statistics package.

Excel seems also the de facto standard for data entry. If used “with discipline”, then data entry can be effective with Excel for most data sets. We explain what we mean by “with discipline” in one of the “good-practice” guides, called:

“Disciplined Use of Spreadsheets for Data Entry”

There is also a second guide to explain what types of data set would be too complicated for entry to a spreadsheet. This is called

“The Role of a Database Package in Managing Research Data”

There is software to support data entry with checking. One possibility is through the statistics package, SPSS, which is an exception in statistics packages and has a special data entry module. An alternative is EpiInfo, which is free to download from http://www.cdc.gov/epinfo.

Statistics packages differ in the formats that they can import, but Excel is supported by all. This is another reason why keeping data in Excel files is sensible.

Most statistics packages offer the Open Data Base Connectivity (ODBC) features, described for Instat in Section 7.6. This allows the import of data from databases and other software, choosing first the tables of data and then the variables and records to import by specifying the appropriate query.

In Section 7.10 we described Instat’s facilities for coping with missing values. Missing values require careful handling and all standard statistics packages have similar facilities. If you are using Excel, then our SSC-Stat addin always takes a blank cell as a missing value. This is also the case for most other routines within Excel, but some seem to replace a blank cell by zero, so tread carefully.
8.1 Introduction

In the next two chapters we show how the data in a worksheet can be organised, where necessary, prior to an analysis. Here we consider simple transformations that leave the shape of the data unaltered. In the next chapter we consider transformations, such as selecting subsets, that change the shape.

Most transformations use Instat’s Manage menu, as shown in Fig. 8.1a. The previous chapters were largely concerned with the Manage ⇒ Data sub-menu. In Section 8.2 we move down the Manage menu to consider Manage ⇒ Calculations and Manage ⇒ Transformations ⇒ One to One. The other options on the Manage ⇒ Transformations sub-menu, shown in Fig. 8.1a, are used in Chapter 9.

Fig. 8.1a Manage ⇒ Transformation menu

In the remaining sections of this Chapter we look at examples where the individual measurements need to be summarised, before the data are analysed. These summaries are often called “indices” or “indicators” and we give examples in Sections 8.3, 8.4 and 8.5.

A common example is often called “repeated measures” data, a topic that is considered in Chapter 19. For example, data on a set of children might include general information in x1-x6 together with the marks on twelve tests in x7 to x18. We might wish to summarise each child’s performance by giving the mean of the twelve marks. Perhaps also the standard deviation of the marks would reflect the extent to which a child is consistent. These summaries are then the “indicators” of the child’s performance.

Whether the calculation of these indices is considered as the first stage in the analysis, or as organising the data is perhaps playing with words, but it does relate to the menu options that we use. The calculations are often simplest using the Statistics ⇒ Data Summary ⇒ Row Statistics dialogue, where the menu is shown in Fig. 8.1b, though the Manage ⇒ Calculations dialogue can alternatively be used.

Continuing our example of the 12 columns of children’s marks, we show how to calculate the ordinary mean in Section 8.3. If some of the subjects, i.e. some of the columns x7 to x18, are more important than others, you may wish to calculate a weighted mean instead, and this is described in Section 8.4.

In Section 8.5 we relate the index we have calculated, to a “standard value”. This is the subject of “index numbers” that is common in business, for example for price or inflation indices, but the idea is general.

In Section 8.6 we look briefly at the calculation of some diversity indices that are commonly used in ecology.
8.2 Calculations

We illustrate with the worksheet for simple regression, introduced in Chapter 3. In Fig. 8.2a we show the Manage ⇒ Calculations dialogue, where we will produce a new column containing the square of the values in X1. Fig. 8.2b shows the result. This is a good illustration of the way a statistics package is a column calculator.

Instat provides a variety of ways of doing transformations. Calculating polynomials can also be done using the Manage ⇒ Data ⇒ Polynomials dialogue.

Common transformations can also be done using the Manage ⇒ Transformations ⇒ One to One dialogue, shown in Fig. 8.2c. This dialogue requires the user to consider the type of data they have. Fig. 8.2c shows the options for non-negative data and the power transformation, that we used above, is one example.

With the transformation dialogue in Fig. 8.2c, Instat can also give the inverse for some of the transformations. This can be useful after an analysis, so the results are displayed on the original scale.

A second option of the Manage ⇒ Transformations ⇒ One to One dialogue is shown in Fig. 8.2d. Here we have used the rice survey data. The dialogue shows how the fertiliser
amounts that were in column X4 can be recoded to a 3-point scale, where we have chosen to have zero, “less than 2” and “2 or more” as the three categories. The resulting column could then be declared as a factor, as we showed in Chapter 7, before being used in the analysis.

To illustrate the flexibility of the calculation facilities, we show, in Fig. 8.2e, how this same recoding could be done using the Manage ⇒ Calculations dialogue. This makes use of the logical functions and the corresponding formula is

\[ X8 = 3 + (X4 < 2) + (X4 < 0.1) \]

When the terms in brackets are FALSE, they return zero, and when they are TRUE, they give –1. Thus, X8=3 when X4>=2. It takes the value 2 for X4 between 0.1 and 2 and otherwise X8 = 1.

?

Finally, in Fig. 8.2f, we show how summary statistics, such as the mean, can be used in formulae. The example shown, uses the formula

\[ x9 = \frac{(x6-mea(x6))}{sde(x6)} \]

This has “standardised” the column, and x9 has mean zero and standard deviation of one. If you are worried that you might forget this formula, then it is always available as an option in the Manage ⇒ Transformations ⇒ One to One dialogue, see Fig. 8.2d.

8.3 Calculating with rows

The fundamental data structure in a statistics package is the column of data. Most of the dialogues that manipulate or perform calculations on data, operate on columns. However, situations do arise in which we would like to have summary statistics of rows of data as well as columns.

If we had wanted to perform operations only on rows and not on columns, then we would have entered the data into the worksheet the other way around. So we are considering those situations where most of our data manipulation will operate on columns, but we would also like to have some facilities for working with rows.

This can always be done, using the Manage ⇒ Calculate dialogue, but it is often easier with the Statistics ⇒ Summary ⇒ Row Statistics dialogue, shown in Fig. 8.3a.
The example below uses the monthly rainfall data from Chapter 3. There, the data for each month is in a different column and we calculate the annual totals in X14, using the Row Statistics dialogue in \textit{Fig. 8.3a}. The result is shown in \textit{Fig. 8.3b}.

\textbf{Fig. 8.3a} Statistics $\Rightarrow$ Summary $\Rightarrow$ Row Statistics \hspace{1cm} \textbf{Fig. 8.3b} Results

This same calculation, with the Manage $\Rightarrow$ Calculate dialogue, would have been as follows:

\[ x_{14} = x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13}. \]

\section*{8.4 Calculating indices}

The poverty survey, introduced in Chapter 3, is shown again in \textit{Fig. 8.4a}. This typifies a study where we need to calculate an index, that will then be used in the analysis. In \textit{Fig. 8.4a} the data in column x3 show which family received the “starter-pack”, to help in their cultivation. This was intended for the poorer families and the main aim of the survey was to evaluate whether the support had really been targeted for these families.
In the study there was information on both the “assets”, like the number of chickens, and the “income” of each family. It was decided to calculate two indices, one of assets and the other of income. These are shown in $x_{26}$-$x_{29}$ in Fig. 8.4b. These were then combined to form a poverty index, given in $x_{30}$, in Fig. 8.4b.

Often, as here, indices are weighted means or sums of the columns of data. As such they can be found using the Manage $\Rightarrow$ Calculate dialogue, or an extension of the Statistics $\Rightarrow$ Summary $\Rightarrow$ Row statistics dialogue that was introduced in Section 8.3.

The weights to calculate the assets index are shown in Fig. 8.4c. They were found from a separate participatory study that included detailed discussions and data collection from groups of villagers.

### Fig. 8.4c Weights for asset index

<table>
<thead>
<tr>
<th>Column</th>
<th>Asset</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>X5</td>
<td>Chicken</td>
<td>2</td>
</tr>
<tr>
<td>X6</td>
<td>Doves</td>
<td>2</td>
</tr>
<tr>
<td>X7</td>
<td>Ducks</td>
<td>2</td>
</tr>
<tr>
<td>X8</td>
<td>Guinea Fowls</td>
<td>2</td>
</tr>
<tr>
<td>X9</td>
<td>Pigs</td>
<td>15</td>
</tr>
<tr>
<td>X10</td>
<td>Goats</td>
<td>10</td>
</tr>
<tr>
<td>X11</td>
<td>Cattle</td>
<td>100</td>
</tr>
<tr>
<td>X12</td>
<td>Bicycle ownership</td>
<td>20</td>
</tr>
<tr>
<td>X13</td>
<td>Radio ownership</td>
<td>7</td>
</tr>
<tr>
<td>X14</td>
<td>Oxcart ownership</td>
<td>100</td>
</tr>
</tbody>
</table>

The Statistics $\Rightarrow$ Summary Data $\Rightarrow$ Row Statistics dialogue, as shown in Fig. 8.4d can be used to calculate this index. An alternative is the Manage $\Rightarrow$ Calculate dialogue, or command as

$$X_{31}= 2^* (x_5+x_6+x_7+x_8)+15^*x_9+10^*x_{10}+100^*x_{11}+20^*x_{12}+7^*x_{13}+100^*x_{14}$$
The assets were divided into 4 categories. We therefore used the Manage ⇒ Transformations ⇒ One to One dialogue, to recode the data, as shown in Fig. 8.4e. The results are shown in Fig 8.4f, where \( x_{31} \) and \( x_{32} \) are to be compared with the original calculations of \( x_{26} \) and \( x_{27} \). The final step is to use the Manage ⇒ Data ⇒ Set factor dialogue, shown in Fig. 8.4g, to add the labels to \( x_{32} \).

Further details of the calculations of the income index and their combination to produce the poverty index, are in the main report of the Monitoring & Evaluation Programme, Sarah Levy, (Calibre Consultants) and Carlos Barahona (Statistics Services Centre, University of Reading) September 2001. This report is available from the Statistical Services Centre (statistics@reading.ac.uk). The resulting poverty index was in five categories, ranging from
“poorest” to “least poor”. These labels are intentional, because no one could be considered “rich”.

8.5 More on index numbers

To some readers to phrase “index numbers” will remind them of the “price index”, or “inflation index”, or “stock-market index”, that is often publicised in the news. These indices are simply weighted averages, or weighted averages calculated as described in Section 8.4, but then standardised so that the base value was 100.

We illustrate with a simple example, using the data shown in Fig. 8.5a. There we consider a price index for running a car. The four columns of data in x2 to x5 are the price of a particular car, the cost of petrol, the cost of servicing or repairs and the road tax. We then multiplied by the amounts we use each of these components in a year, and took the values 0.2 for x2, to correspond to buying a car every 5 years. The 4 values we gave were 0.2, 2000, 5, 1, as shown in the use of the Statistics ⇒ Data Summary ⇒ Row Statistics dialogue, shown in Fig. 8.5b.

Often these index numbers are for “monitoring data”, collected regularly over time, and the monthly index of inflation is an example.

However, the idea is more general. The data could equally be collected at different sites in the country, or could be income for different people in a firm. As a second example we again use the Malawi poverty survey, described in Chapter 3, and Section 8.4. There the income index had already been calculated and is shown in x28 in Fig. 8.5c. We “standardise” by considering 550 Malawi Kwacha (about $7 or £5) as a baseline monthly income. Simple use of the Manage ⇒ Calculate dialogue, shown in Fig. 8.5d, is all that is needed to calculate the (standardised) index.
8.6 Diversity indices

Indices do not have to be weighted averages. We illustrate with the calculation of diversity indices, that are commonly use in ecological studies, though no special knowledge of ecology is needed for this section. As an example we use the coral reef data, described in Chapter 3. Some of the data from one site in that study are shown in Fig. 8.6a. The data shown are the abundance, i.e. the number of times each type of coral was observed in each of 4 transects.
We consider first the data in transect A, i.e. in x3. The simplest index is sometimes called “Richness” and is just the number of different species in the transect. That is just the number of non-zero values in the column and we see by inspection, in Fig. 8.6a, that there are 8 such values in x3. We ignore for now that the last 3 rows are dead coral, rubble and sand, and at least the last two these are not really species of coral!

Fig. 8.6c Statistics ⇒ Summary ⇒ Row

The layout of the data in Fig. 8.6a may make this calculation seem different to the simple indices calculated in Sections 8.4 and 8.5. To emphasise the connection, we have transposed the data from Fig. 8.6a to the layout shown in Fig. 8.6b. Now we have used our familiar Statistics ⇒ Data Summary ⇒ Row Statistics dialogue, as shown in Fig. 8.6c, to get the count of non-zero values. The results, for all four transects, are shown in the last column in Fig. 8.6d. This time we have omitted dead coral, rubble and sand in calculating the index.

The results in X21, shown in Fig. 8.6d, are that there are 7, 6, 7 and 4 different species of coral in the four transects, that were sampled. What is also of interest is to find how many different species of coral there were overall. The answer is not necessarily 7+6+7+4 = 24, because the same type of coral was observed in the different transects. Indeed we see that this is the case, by looking at the data in Fig. 8.6a. There we see that 14 different types of coral were observed in the 4 transects. We show below how these data can be used to estimate how many different types of coral there might be, had we looked at the whole part of the reef and not just a sample of four transects.

In Fig. 8.6e, we introduce a special dialogue that can calculate the richness and some other diversity indices that are commonly used in ecological studies. We use the data in the “column” form shown in Fig. 8.6a, because that layout is more common in ecological studies.
The results for the analysis of a single column, transect D, are shown in Fig. 8.6f, and we see that the “Richness”, that we calculated earlier, is the first index. The indices are all quite simple formulae that are given in the Reference Guide and could be evaluated with a macro, if no dialogue exists. As an example, perhaps the most complicated formula is the Shannon index of diversity and this is commonly used. For transect D is given by ignoring the zeros (which they all do) and calculating:

\[ H = -\sum \left( \frac{n_i}{n} \ln \left( \frac{n_i}{n} \right) \right) \]

\[ = - \left( \frac{11}{32} \ln \left( \frac{11}{32} \right) + \ldots + \frac{1}{32} \ln \left( \frac{1}{32} \right) \right) = 1.48 \text{ in Fig. 8.6f.} \]

The “ln” in the formula is the natural logarithm.

In Fig. 8.6g we show the Statistics ⇒ Data Summary ⇒ Diversity dialogue again, but this time we look at all four transects. We have to specify the layout of our data, and in Fig. 8.6a, each column is a sample (i.e. a transect) and each species is a row. This is the second of four options in Fig. 8.6g.
We choose the simple “Richness” index again and the results are shown in Fig. 8.6h. The summary values are slightly larger than shown in Fig. 8.6c, because we are back to including dead coral, sand and rubble as special “species” again. We have used an option in the dialogue in Fig. 8.7g, to save the results and this puts the 4 values for the index into a column (just like the Row dialogue) so the index can be used in further work.

As a pointer to the methods of inference introduced in Chapter 15, we have included the Jack-knife option to estimate the total number of species in this part of the reef, in the dialogue in Fig. 8.6g. The estimate is 25 species, but with wide confidence limits, partly because we only have a sample of four transects.

The calculation of these indices is included here because the ideas are the same as those in the previous sections. We continue with this example in the next Chapter. There we look particularly at the way the structure of the data causes complications in the analysis of studies of this type.

8.7 In conclusion

All other statistics packages have an equivalent facility to the calculate dialogue in Instat. In some it is more powerful, and also allows calculations on other structures such as matrices.

Like many packages, Instat can be used as a calculator and this can be of more general use than statistics. This may be a useful feature for some, who then start using a statistics package as a routine.

We start by File ⇒ New Worksheet, which also provides “memories” for the calculator. Then either use the Manage ⇒ Calculate dialogue, as shown in Section 8.2, or simply type into the Commands and Output window, as shown in Fig. 8.7a.

If the calculations are typed, then put a “?” before the calculation to display the result, as in:

: ? 2 + 3 + 4

Some examples are shown in Fig. 8.7a. There we also have the constants, k1, k2, etc. that are part of each worksheet and can be used as memories to store single values.
The Commands and Output window provide a scrollable record of the calculations that were done.

**Fig. 8.7a Calculating with constants**

![Image](image1.png)

We show in **Fig. 8.7b**, how the columns, x1, x2, etc. can also be used. They therefore provide “memories” to store whole columns of information. The calculations in **Fig. 8.7b** also demonstrate how a statistics package is a “column calculator”. For example we have entered x1 and then stated

: x2 = x1/25.4

: ?sum(x2)

**Fig. 8.7b Calculating with columns**

![Image](image2.png)

Of course, all these calculations can also be done from the comfort of the Manage ⇒ Calculations dialogue, as shown in **Fig. 8.7c**. The results are in **Fig. 8.7d**.
The result of that calculation is shown (in bold and blue) in Fig. 8.7d. Either way, the output window provides a record of the calculations as well as the results.

In Sections 8.4 to 8.6 we considered the calculation of various indices. Most general statistics packages would use the general calculation facilities for such tasks, rather than the special dialogues that have been used here. For more specialised indices, such as the diversity, some special software is also available. The web site www.pisces_conservations.com gives some examples.
Chapter 9 - Changing the shape of the data

9.1 Introduction

In Chapter 8 we considered dialogues that transform the data, but do not change its shape. Here, in Section 9.2, we see how short columns can be joined, or stacked, together. In Sections 9.3 and 9.4 we see how a subset of a column can be selected. In Sections 9.5 and 9.6 we show how data can be transposed, so the rows become columns, and how columns of data at multiple levels can be merged.

When datasets have a complicated structure, users sometimes have to spend most of the time on data manipulation. In Sections 9.7 and 9.8 we consider how to spot potential problems and hence how to minimise the time that has to be spent sorting them out.

9.2 Stacking data

For illustration we use the experimental data from Chapter 3, Section 3.4, which we show again in Fig. 9.2a. However, we assume that the data were, instead, entered in the form shown in Fig. 9.2b. These data are in the worksheet called "experi2.wor", to distinguish them from the data in the file called "experi.wor", that were entered in the usual form for analysis.

We often find students arrange the data as shown in Fig. 9.2b especially if they are using a spreadsheet to enter the data. This arrangement is often shown in a textbook, because it is a convenient layout to illustrate a "hand" analysis of the data.

For the computer solution, we require the three columns of data to be "stacked" as one column, with a second factor column to indicate which column the data originally came from.

One temptation is then to quickly retype the data. We urge users to resist this temptation, as it may easily lead to errors. Instead use the computer to re-organise the data. The appropriate dialogue is accessed using Manage ⇒ Transformations ⇒ Stack and is shown in Fig. 9.2c.
Note that as well as joining the columns, we have also used the dialogue to add x5. This is a factor column, that specifies the original columns, corresponding to the three treatments.

To complete the task, see Fig. 9.2a, we can add a label column for the factor and we still need to add a further column to specify the blocks. This can be typed, or we can use the Manage ⇒ Data ⇒ Regular Sequence dialogue.

Sometimes, when stacking data you may find that the maximum length of the columns in the worksheet is not long enough. Then use Manage ⇒ Resize Worksheet first, to increase the number of rows.

Occasionally, we need the reverse operation, which we call Unstack. This is not shown, but uses a similar dialogue, Manage ⇒ Transformations ⇒ Unstack, to split long columns, such as x4 in Fig. 9.2d, into the short columns, like x1-x3.

9.3 Selecting subsets

The Manage ⇒ Transformations ⇒ Select dialogue is used to select subsets of the rows. For illustration we use the 32 years of monthly rainfall data described in Chapter 3 and shown again in Fig. 9.3a.

The years 1956 and 1957 were missing so as an example, we choose the years from 1958 onwards, as shown in Fig. 9.3b. In the “condition for selection” we have used a special variable called “row” that is recognised by Instat. The formula was given as “row>=7”. We could alternatively have given the condition as “X1 >1957”. 
Often the criterion for selection is based on the levels of a factor. This may be specified in the formula, but if a single factor is used, then a simpler alternative is shown in Fig. 9.3c. There we have used the survey data, shown first in Chapter 3, and have selected just the first 2 villages. In Fig. 9.3d we have also shown part of the printout from the output window. This indicates that 17 rows of data were selected, and that the columns in the original data that were factors have been set to be factors in the subset as well.

Sometimes you may not have enough columns for the full set of data and the subset as well. Then one solution is to resize the worksheet. Alternatively, the subset can be saved into a second worksheet.

One way to save the subset into another worksheet is to use File ➔ Save As first, to save a second copy of the data. This copy is then automatically opened, so the original is safe. Then, as indicated in Fig. 9.3e, you can specify that the “Select” dialogue writes into the
same columns as the original. It will then overwrite those data. If warnings are in force, then
you will be asked if you want to overwrite the columns, as shown in Fig. 9.3f. Answer “Yes”,
or use Edit ⇒ Flags ⇒ Disable Warnings first.

9.4 Selecting and Stacking

Occasionally we need to select data from multiple columns and put the results in a single
column.

The example we use here consists of the daily rainfall data from 11 years in Northern
Nigeria. The file is called samsmall.wor, and part is shown in Fig. 9.4a. Here each of x1-
x11 contains one year of data. We have added x12 with the numbers from 1 to 366, giving
the day-number of the year. (Why a year always has 366 days is explained in the Instat
Climatic Guide.)
The task we suppose is to select the first occasion after 1\textsuperscript{st} April, i.e. day 92 in the year, when there is more than 20mm of rain. This could define the start of the rainy season. We would like the amounts in x13 and the day numbers in x14.

**Fig. 9.4c** Manage $\Rightarrow$ Transformation $\Rightarrow$ Select and Stack  **Fig. 9.4d** Results

Here we use **Manage** $\Rightarrow$ **Transformations** $\Rightarrow$ **Select and Stack**, as shown in **Fig. 9.4c**. It is called “select and stack” because it takes the selections from multiple columns and stacks them into a single column.

In the dialogue we have opted to select just the first time that this condition is satisfied.

Another new feature in **Fig. 9.4c** is that the condition is given as

\[ X12 \geq 92 \text{ and } \text{col} \geq 20 \]

The term “\text{col}” in the formula is recognised by Instat and is interpreted as representing each of the input columns, in turn. The condition is therefore evaluated for x1, x2 to x11 in turn, giving columns of 11 observations. The results are shown in **Fig. 9.4d**. For example in the 7\textsuperscript{th} year (as shown in the data columns in Fig. 9.4a), the first occasion was on day 113, when there was 30.48mm.

The results can be displayed more clearly using the **Manage** $\Rightarrow$ **View Data** $\Rightarrow$ **Format** dialogue, shown in **Fig. 9.4e**. This includes a special “Day of year” format, and we see that the starting day ranges from 17\textsuperscript{th} April to the 4\textsuperscript{th} of June.
9.5 Making rows into columns

Statistics packages are “column calculators”, unlike spreadsheets, where the data can be orientated as we like. So sometimes we need to transpose the data.

For illustration we use the monthly rainfall data, from Chapter 3, Section 3.5. We begin by removing the extra columns that may have been generated from the work in Section 9.3. Use either the Manage ⇒ Remove dialogue, as shown in Fig. 9.5a, or type

: Remove x15-x40

in the Commands/Output window.

One way of transposing uses a special case of the Manage ⇒ Transformations ⇒ Select dialogue that we used, in Section 9.3, to choose subsets of the data. It is shown in Fig. 9.5b and 9.5c. In Fig. 9.5b we have the dialogue ready to transpose the whole data, while in Fig.
9.5c we selected the data for a single year. In the second case, in Fig. 9.5c, we have to know how many rows will be selected, and complete the “Into” box ourselves. Here we know that the condition

\[ X1 = 1954 \]

will be satisfied by just one row, and with the transpose, this becomes one column. Hence we have put x47 into the “Into” box.

**Fig. 9.5c Transposing one row**

**Fig. 9.5d Transposed data**

We have added a name to the columns in Fig. 9.5d, using the Manage ⇒ Data ⇒ Name dialogue, because it allows a sequence of names to be given. We have also added a factor column, in x48, giving the months. This clarifies the display, and also provides the x-axis for graphs of the data for a given year.

Sometimes it is more convenient to have the 2 sets of data in separate worksheets. This is a 2-stage process, as follows:

- First use File ⇒ New Worksheet to define a new worksheet. Make it one with at least 50 columns, because you need 32 just for the transformed data.

Now use the Manage ⇒ Duplicate (Copy Columns) dialogue, as shown in Fig. 9.5e with the transpose checkbox ticked. In this case we have to know which columns to select from the original worksheet. We also need to know how many rows there are, because Instat does not know at this stage. So in Fig. 9.5c we have entered x1-x32 ourselves.

Part of the resulting data is shown in Fig. 9.5f.
9.6 Data at multiple levels

Many studies involve data at more than one level. We illustrate with a simple example from Chapter 3 that is shown again in Fig. 9.6a. We have a pilot survey of just 8 people where we have recorded their sex and age group. This information, in x1-x3 is the “person” level. Then we have information in x4-x6 about activities. These three columns are at the “activity” level. X5 gives the activity and X6 how long was spent on each activity.

The general idea is that we now have 2 rectangles, instead of one. There is one at each level, together with a column that links the levels. So here x1 gives the person and we also see, in x4, which person did each activity.

For example, we see that person 1 did 3 activities and spent 30, 20 and 60 minutes on each of walking, jogging and cycling. Person 2 did just one, and the absence of information on person 3 indicates that he did none.

Fig. 9.6a Data at person and activity levels
Many studies involve data at more than one level. For example, a survey may have some
information at the household level and then interview each person in the household. In
education we may have information from teachers at the classroom level and from each
pupil at the person level. In an agricultural experiment we may have yield at the plot level
and also height and disease score at the plant level.

We will look at the analysis of the example in Fig. 9.5a in Chapter 13, when we deal with
multiple response data. Here we are concerned simply with the organisation of the data
between these two levels. For example, if we are working at the “activity” level then we may
wish to expand some of the “person” level information down to this level. We use the
Manage ⇒ Transformation ⇒ Expand dialogue, shown in Fig. 9.6b. Fig. 9.6c shows the
results with both sex and ages copied to the “activity level”.

Fig. 9.6b  Manage ⇒ Transformation ⇒ Expand  Fig. 9.6c  Activity-level data

The opposite is to summarise the activity-level data, for an analysis at the person level. In
Instat, one way this is done is with the Statistics ⇒ Summary ⇒ Column Statistics
dialogue, shown in Fig. 9.6d. We see again the information at the person-level in Fig. 9.6e,
and the statistics dialogue has added summary information at this level, giving the maximum
and the total time spent on exercise.
This second task, of summarising the information up one level, is a very common requirement. For example, in a survey we take the total income of all members of the household, to analyse at the household level. In an experiment, we take the average height of the plants, which is then analysed at the plot level.

9.7 Coping with difficult datasets

Sometimes the way the data have been entered makes it difficult to analyse so that all the objectives of the study are satisfied. The difficulty is usually that the analysis has to proceed in stages and the layout of the data is not appropriate for one of these stages.

We use the coral data, described in Chapter 3, Section 3.8 as an example. Some of the data are shown in Fig. 9.7a. However the problem is general so we also explain how a household survey could give data with the same complications.
The coral data were collected from 17 different sites, of which one is shown in Fig. 9.7a. At each site there were 4 transects, called A, B, C and D. The data give the cover and the number of occurrences of each species that was observed at the site.

The structure could be the same in a survey of 4 households in each of 17 villages. In each family the data indicate the reasons that the members of the family used their car. So, the frequencies of going to the doctor, cinema, shops, school, etc, would be the different rows of data.

The layout shown in Fig. 9.7a is fine for the first part of the analysis, which is to calculate some (diversity) statistics separately for each of the 17 sites. The difficulty is that some of the objectives of the study require the data to be compared for the different sites. This is more awkward, because the data for each site is in a separate rectangle.

The main problem occurs because the set of species (or set of reasons in the household survey) is different for each set of data. It will be even worse if there are any spelling mistakes in the names of the species, because the computer will interpret any different spelling as a new species.

There are two alternative strategies. The first is to reorganise the raw data, so they can be used for both stages in the analysis. The second is to leave the raw data as above and just transfer the summary values from the first stage. We look briefly at these options in turn.

In Figs. 9.7b, 9.7c and Fig. 9.7d we show ways that the data could be organised so that both stages of the analysis would be possible.

Fig. 9.7b Coral data for all sites

Fig. 9.7b keeps the data roughly as shown in Fig. 9.7a, but has omitted the summary columns and then combined the 17 sites as successive rows in a single rectangle there are 333 rows of data in total.
Fig. 9.7c  Transposed data, all sites

Fig. 9.7c shows the data from Fig. 9.7b transposed, so that each of the 101 species is a column. There are 68 rows of data, with 4 transects from each of the 17 sites. With this layout each column contains the counts for a particular species.

Now we can easily calculate row statistics (for example indices of diversity), and these could be put into columns for further analysis.

One problem with the layout Fig.9.7c can occur if there are many possible species. Then the dataset becomes very wide, with most of the counts possibly zero. The “list-format” in Fig. 9.7d indexes each count by its transect and species. In Fig. 9.7d we have kept the zeros from Fig. 9.7b, but they could be eliminated with the layout of the data. From Fig. 9.7d, we can do a cross tabulation if we wish to look at the data in either of the forms shown in Fig. 9.7b or Fig. 9.7c.

Fig. 9.7d  Coral data in list format

Fig. 9.7e  Coral data for the second stage

28/06/2002 106
In a case study we explain how to get from the original format to any of the formats shown in Figs. 9.7b, 9.7c or 9.7d. The data were originally in Excel, and this is a convenient package to reorganise the data.

The alternative strategy is a two-stage process. First we analyse the data in each individual transects and then we transfer the summaries so that the second stage can be done. For example, Fig. 9.7e shows the data for all 17 sites after the first stage has been done. Here there are columns representing two diversity indices.

This second strategy is often done, because it seems so much simpler. But it has two limitations. The first is that the analysis for any new index often takes some time, and hence we find that the data are incompletely analysed. The second is that some summaries are more difficult. For example we might later decide we would like to look at measures of similarity between the different sites. There are many definitions of similarity, but they would each depend on looking at the data from different sites. This is simple once the original data are reorganised, but not if they are kept separate, as in Fig. 9.7a.

9.8 In conclusion

All the main statistics packages have similar facilities to those shown here for stacking and unstacking data, and for selecting subsets. Most also have facilities for merging data that are at multiple levels that are more powerful than those in Instat. However, with SAS as an exception, statistics packages are built mainly to analyse rather than to manage data. So if you have data at multiple levels, then we suggest you consider managing the data in a database package. You can then use ODBC or other importing facilities to import the data you need for particular analyses.

If you use a spreadsheet, then you are likely also to need to be able to stack and un-stack your data. Excel does not provide this facility and a flexible stack and unstack is therefore one feature of our SSC-Stat add-in.

Excel provides the VLOOKUP function to link multiple level information. SSC-Stat includes a dialogue called LOOKUP to make this function easier to use.

Sometimes the requirements for data organisation are sufficiently complicated that the quickest solution is to write a special macro. The coral data, used in Section 9.7 is an example. There we had 17 "rectangles", in a single Excel sheet, corresponding to the 17 sites used in the survey. Within each site there is a row of data for each type of coral, and the complication was that the set of coral types was different at the different sites.

Our task, for the reasons described in Section 9.7, was to combine the data into a single rectangle. We could not see how to do this, using Excel's standard functions, nor using the extra facilities programmed into SSC-Stat. So we wrote some simple macros. The results, i.e. the transposed data, were shown in Figs. 9.7b, 9.7c and 9.7d. For those who would like to see the macros, they are within the Excel workbook for the coral data.
Chapter 10 - Developing a strategy

10.1 Introduction

We have now reached the end of the chapters devoted to organising the data. The remainder of this guide is concerned with the facilities for analysis. Before embarking on these chapters we reflect on the types of strategy that users might adopt, if they are to process their data effectively.

It is more important that you have a strategy than exactly what it is. Here we consider both “you” as an individual and your “group”. The “group” might be your section, or department, or faculty, or organisation.

To give examples, your strategy might be to use Excel for everything. Or to use any other single package, like Mstat, or SPSS, in the same way.

An alternative strategy would be to use Excel for data entry, with a database package if your data are complex. Then you use a statistics package for the analysis and a special graphics package for drawing the graphs. You just use a simple statistics package, like Instat, but there is a more powerful statistics package that you use, with help from colleagues, when you need it.

We find some people are easily frustrated by the differences between packages and would prefer to use a single statistics package, "warts and all". Others enjoy using a mix of statistics packages. For them, the choice of the right package for a given task, can both enhance and speed their analyses.

Cost of software is an issue, but for many packages the cost of the statistical software is modest. For example, in a University in a developing country, we usually hear there is just no money for software. However donors have supplied computers, staff members are financed to travel and postgraduate students are supported for training overseas. If organisations have a strategy, then donors we have approached would be just as happy to supply software. Powerful statistical packages could be supplied for the whole faculty for the cost of one visit abroad. It is largely a question of priorities.

More important is to consider the “full costs” and the initial cost of buying the software may be a small component. Often much more is the “time-cost” of learning to use the packages within the strategy that is adopted. Then, in a training environment, there is the hidden cost of poorer or more difficult training for staff and students, if the appropriate software is not available. Research and development projects often include large costs of collecting data, and this can be wasted if the data are then poorly or incompletely analysed and reported.

In an evaluation of statistical aspects of research projects in 2002 our Centre found that some researchers reported their results in exactly the way they were produced by the software, so poor output by the statistical software resulted in poor presentation in the reports.

In this chapter we consider a possible strategy for (statistical) software in Section 10.2 and then look at four areas where the software is used, in the following sections. These areas are data organisation, data analysis, reporting and also training. We have tried to be generic. So we suggest how readers could develop their own strategy, rather than prescribe one for them.
10.2 Software

Most statistical packages try to be comprehensive in the facilities they offer to support users in their statistical work. It has to be remembered that some of the current statistical packages began in the 1960’s. This aim of being comprehensive was very useful until the 1990’s, because data exchange was not easy and learning to use each package took some time. Now however, many users have opted for a mix of packages. For some this has been a conscious move, while others simply found that a spreadsheet environment, particularly using Excel, was easier than using a statistics package.

Some statisticians have attempted to counter this use of spreadsheets by pointing out the deficiencies of some of the algorithms used for the statistical analyses. Those criticisms are correct, but they apply largely to those who use a spreadsheet for all their work.

Instead we would like to encourage users to consider a strategy that allows for an appropriate mix of the software they need for their statistical work. One assumption we make is that users may need more than a spreadsheet. That might not be the case if only descriptive methods are needed for the analysis. These are the methods described in Chapter 11 to 13 of this guide, namely graphs, and summary statistics including (pivot) tables.

Many people start from the basis of being very comfortable with a spreadsheet. By this we mean that you already use it for data entry and data manipulation. If this is your situation, then the key question is whether your data processing would be improved by adding one or more statistics packages. The next chapters may help, in that we view spreadsheets as largely adequate for descriptive statistics, i.e. up to the end of Chapter 13 in this guide. If you need the materials in the later chapters, particularly from 16 onwards, then we suggest a statistics package will be useful.

Those who add a statistics package will usually continue to use a spreadsheet for data entry and initial summary. Spreadsheets are also good in drawing graphs and the results might therefore be transferred back from a statistics package for graphical presentations.

Although most users know something of spreadsheets, you may not yet use it much for statistical work. For you, there are two alternative strategies. The first is to learn to use the spreadsheet for the basic statistical work, and the a statistics package for the “analyses a spreadsheet cannot reach”. Or perhaps you could just learn to use a statistics package for all your analyses.

Your decision will depend on the level of statistics you need and how long it takes you to learn new packages. The statistical software can handle all the simple data summaries, i.e. up to Chapter 13 in this guide, but so can spreadsheets, if that is all you need. For more statistics you should use a statistics package, so your ease at using a mix of packages is a factor.

10.3 Organising the data

We guess that some people will skip this section, because they are not interested in data management. However, we have found that analyses are frequently hampered by poor data organisation. It is an area where initial thought, and a well-defined strategy, will pay huge dividends in all but the smallest projects.

We start again by those who currently enter and organise their data in Excel. There is no need to change, but you may wish to check on how effectively you are using Excel for the data entry. Briefly we suggest that you make use of what Excel calls “list format”, you should enter each “level” of your data into a separate sheet, and you should not mix data entry with data summary. If you need more information then our guide titled
“Disciplined Use of Spreadsheets for Data Entry”

may help.

Sometimes your data may be too “complicated” for a spreadsheet. Then you might want to consider using a database package, such as Microsoft Access, for the data entry and management. We have a guide called

“The Role of a Database Package in Managing Research Data”

In that guide we explain what we mean above by “too complicated” and also describe alternative strategies for those who start to use a database package. One option for those who have complex data is to investigate the package called Epilinfo. This is designed particularly for surveys on health issues, but the software is general, and is free. It also includes special facilities for including data checks, and to support “double-entry” of data. This feature is lacking in Excel and in most database packages.

An alternative is to enter your data directly into a statistics package. If you do not need data checking facilities, then any can be used. Otherwise your choice could include SPSS, that has an special module for data entry, Genstat for its support of double entry and SAS, that is built to handle very large volumes of data effectively.

10.4 Data analysis

We start by looking generally at some of the different statistics packages. In the “In Conclusion” section of the individual analysis Chapters we also mention points about some of these packages for particular types of analysis.

The package called SAS (Statistical Analysis System) has become effectively an industry-standard among statistics packages. This is similar to the dominance of Excel among spreadsheet packages though not to the same extent (at least when this was written – in early 2002). So SAS is the only package that could make a good claim to be in our “in conclusion” section in all chapters.

The other dominant package is SPSS (Statistics Package for the Social Sciences). This is particularly well known for the analysis of survey data, see Chapter 13.

If you already use either of these “giants”, then, just as with Excel, the issue is usually not to change from them, but whether the addition of any of the other statistics packages would help your strategy for analysis.

If you do not currently use a statistics package, or if you would like to add to your use of SAS or SPSS, then there are still a good many that are vying for your custom. Contenders among the general packages include Genstat, Minitab, R, S-Plus, Stata, Statistica, and Systat.

We stated in the introduction that Instat could be a stepping-stone to help users to decide on their strategy for analysis. You may have agreed with us, if you are reading this section. Until a few years ago, we claimed that one reason for using a stepping-stone was that packages like Instat and Minitab were easier to use than those that included the more complicated methods. This is no longer the case, as all are reasonably easy to use. So reasons now for using Instat as a stepping-stone are to help you step more effectively and in the most suitable directions.

Of course the main reason for Instat, is the same as for any of the other statistics packages, and that is that it is useful for at least some of your needs.

Here, we do not attempt a direct comparison of the packages. Instead we look at your possible needs, and then show how these can be used to help in your choice of package. In this assessment consider also the requirements of your “group”, e.g. the University or organisation. For example you may usually find a simple package to be sufficient, but
occasionally need more advanced methods. Hence if a powerful package is available somewhere in your group, possibly with support, this might be the best way to satisfy your needs.

The use of software that is related to particular types of application is described in the “in conclusion” section of the remaining chapters of this guide. Here we take a more general look. The typical problems we find in data analysis include the following:

- Analyses may be too “simple”. Sometimes simple tables and graphs are all that are provided, in situations where we feel more can be learned from the data, if users were bolder in their analysis.

- Where models are fitted, simple significance tests are sometimes overused. Usually, as we explain in later chapters, a significance test is just the prelude to the interesting stages in the analysis. Often, perhaps because they have been over-emphasised in training courses, significance tests are treated as the main tool instead.

- The structure of the data is not explained or is ignored in the analysis. Data are often at different “levels” as we described in Chapter 9, Sections 9.6 to 9.8. Ignoring this, and other aspects of the structure can give misleading results.

- Particular methods become popular and are then used in an unthinking way. Results may then be presented attractively, perhaps using dendrograms, or other nice sounding displays. But the final results may not satisfy what we call the “so what” test!

- A report emphasises that “all analyses were done using the statistics package called xxx.” The report is then largely printout from the package, and it is clear that facilities in the statistics package, rather than the objectives of the study have dominated the analysis.

It is always easier to say what to avoid, than what to recommend. We have made positive suggestions in our “good-practice” guides, that are available within Instat’s help and from our web site. One guide concerns the analysis of survey-type data and is called:

“Approaches to the Analysis of Survey Data”.

A second guide is on the analysis of the sort of data that arises from experiments and is called:

“Modern Approaches to the Analysis of Experimental Data”

Instat can be used for the analysis, using any of the methods described in these guides. Then there is a further guide called:

“Modern Methods of Analysis”.

This 20-page guide mentions some of the major recent advances that we feel are particularly important in supporting effective data analysis, whatever the type of data. In this guide we describe modern graphical methods and also practical uses of multivariate methods. We then consider “generalised linear models” that permit the same methods (like regression) that have been used with the normal distribution, to be applied to other types of data. Finally we describe modern approaches when data are available at multiple levels (like classroom and pupil).

If you need these methods, then you need a powerful statistics package. For some of these methods you would need more powerful software than Instat.

We believe that these methods are sufficiently important that most users should know of their existence and be able to evaluate if they are needed for a particular analysis. If so,
then this part of the analysis might have to be undertaken jointly with someone who knows the methods and software reasonably well.

Although Instat does not include these methods, they are based on the concepts and methods that are in Instat. So, you may still find that Instat is useful in its teaching role, and that is the subject of the Section 10.6.

10.5 Reporting the results

Our guide called “Informative presentation of Tables, Graphs and Statistics” gives our views on the presentation of results. This is mainly designed for written reports, rather than seminars.

Here we consider briefly the more technical side of deciding how you will transfer your results into the form that they can be presented in a report. For example, our strategy, in writing this guide, was to prepare the results in Instat. We usually captured tables and graphs using software called Paintshop Pro. This suited our purpose, because we wanted mainly to show screen shots. We still had to decide on the format in which to save the screen shots, because we also wanted the resulting text to be usable as a Help file.

We have written this guide in Word, and one further decision was whether to import or to link the tables and graphs. We choose to link them, so it is easy to change figures and the resulting Word files do not get too large.

You may feel that your statistics package is fine to produce the graphs in just the way you need them in your report. An alternative is to ensure that your statistics package is capable of easily saving the numerical results, that make the graph. Then transfer the columns to another package that you use for the graphs. Many people use Excel. Then the graph from Excel, or a special graphics package, is transferred to your report.

If you have many similar graphs to include, then make sure you use a graphics package that can easily produce the format that you require. One way is to devise your own “template” for the graphs you wish to produce. For example Excel provides good facilities for users to construct their own templates. Otherwise you will often find that you spend a long time on trivial routine editing of the graphs. More seriously the different graphs may well each be formatted slightly differently, and be hard to compare. For example, graphs of percentages may not all go from 0 – 100%.

Most reports also include tables. Some will be in the text and others may be in appendices. In some projects the raw data are also included. Usually you should use the statistics, or spreadsheet package to produce the tables and then transfer them to your reporting package, such as Word. Reporting software often has excellent facilities for table layout, so you could then edit the labels and other text that surrounds the data in your tables. But do not use the reporting package to type or edit the actual numbers. This is both time-consuming and error-prone.
10.6 Teaching and learning statistics

Statistics is taught at many levels. Here, as an example, we consider a service course in statistics to college students who are non-statisticians, but need it for a research project. Typically they may be starting an MSc or PhD programme, though the same ideas apply to undergraduate students and others.

When (if!) students start their project, one of our initial questions concerns the objectives of their research. Unless these are stated clearly it is difficult for the research to be effective.

As trainers (or trainees) we perhaps ask the same question at the start of our course. The objectives of a service course in statistics should perhaps relate to the students being able to conduct statistical aspects of their research project effectively. Hence we must teach the key skills so this can be done. In many courses this clearly has not happened.

If that is the case, then perhaps we need to make changes. Of course it could be that our statistics training “project” was too ambitious. In a recent study of training needs, the postgraduate students themselves pleaded for changes to undergraduate statistics courses, because they felt that otherwise our objectives could not be met in the time we were given for the postgraduate training.

One direction of change we often suggest was described in Chapter 5. This is to make the training broader, including more on design. We also need more on how to handle reasonably-sized data sets; the “data-management” component that was described in Chapters 6 to 9 of this guide.

On analysis we sometimes find that (service) courses remain too dominated by the theory, both in what is taught, and in the order that the materials are presented. In pre-computer days the order of the topics was usually in order of mathematical complexity. And this was because students had to use the formulae in their practical classes. Computers have been added, but sometimes they have not led to a revision of the training materials.

One reason that is often given is that students still need to be able to do the analyses by hand, or they cannot understand the results, or the role of the computer. We agree with the sentiment, but often not with the extent to which it is applied. That is because the price paid by students is sometimes too high. First, they often do not understand the key concepts and second, the course has not progressed far enough to support their research.

In the remaining chapters we give numerous examples of ways in which we think that the training in statistics can take advantage of the use of computers. This is in addition to the data management ideas that have already been mentioned and which are often omitted from training courses. On analysis we include the following.

- More time can be spent on descriptive statistics. This can include looking at reasonably large data sets of the type that will be used in the subsequent research projects. In the following chapters descriptive methods is about a third of the material on analysis. This part of the training is conceptually simple, and the methods are important for the processing of research data.

- Training on probability models and simple inference (Chapters 14 and 15 in this guide) can concentrate mainly on the concepts, and not just the formulae. There are many inter-related concepts, e.g. normal distribution, sampling distributions, central limit theorem, standard errors, confidence limits, significance levels and so on. If we teach this part of the training effectively, then students have the foundation they need for the later topics. Otherwise the subject will remain confusing forever.

- In teaching further statistical modelling (Chapter 16 and later) we no longer need to go in the order of mathematical complexity. For example, in Chapter 16, we introduce the concept of “blocking” (stratification in a survey) as a general idea, rather than starting
with the simple randomised complete block design. There is no problem with the latter approach except that students may consider only the simple designs, because that was all they were taught. And they sometimes do not understand the real reasons for blocking, because they have mainly been taught how only to analyse a series of special cases. In general a “model-based” approach may help, where students always consider the structure of the data.

So, if you would like to teach, or learn, in a more imaginative way, then what are the software implications?

Here we would claim that Instat is a good contender. It has been designed to support the teaching of statistics as well as to process data. Most other statistics packages are the other way round. They have been designed to process data, though they can be used to support the ideas above in training courses. Minitab is a possible exception in that it has been designed to support both teaching and data processing.

Your strategy may also depend on the software that will be used for later work. If that is a more powerful package than Instat, you may not want the overhead of introducing a different package for the training part and for the project.

The converse is that if two statistics packages are introduced during the course, then it gives students the confidence to use the most appropriate package for their later work, even if it is not one that has been taught.

Your strategy may already include two packages, if you use Excel for part of the training course. Then you could also include some add-ins to Excel as they are required. We have already mentioned our SSC-Stat, which has components to further strengthen Excel for data manipulation and graphics. It also includes dialogues that support the use of Excel for teaching simple inference (Chapter 15). There are many other add-ins that can support the use of Excel for teaching basic statistics. For teaching purposes, we particularly like Berk and Carey (1999).

In Chapters 13 to 17 and 20 of this guide we show many features of Instat that are designed to support the teaching of statistics. An alternative would be to look for special programs, or perhaps special workbooks in Excel, to teach these concepts. These may be better than Instat, because they have been designed for a particular task. We use Instat, partly because there are so many concepts that we do not want to add the overhead of introducing different software. However, some concepts might be helped by the special software, particularly for those who would like to change they way they teach, but do not want to add Instat to the student’s load.

10.7 In conclusion

We find that many commercial organisations are prescriptive in the software they allow, while educational establishments are less so. This is partly due to the higher cost of software to commercial sites and possibly also to the difficulties (like “herding cats”?!?) of dictating to academic staff.

However, in both cases, we suggest it will be in everyone’s interest if there is an identified strategy for the organisation and also encouragement for staff (and students), with different needs, to be able to have their own appropriate strategy.
We would also encourage organisations to consider the support that will be made available to help the software be used as effectively as possible. We have come a long way from the days when data were sent to the statistician, because they were able to handle the software for the analysis. Now the use of Excel typifies a more egalitarian approach. However, perhaps some nominated users, or a central group, or support over the web, might help the packages be used more efficiently. Then we can take full advantage of the new ease of use of modern software, to provide more effective training and analysis.
Part 3 – Basic analysis

We suggest that many users would find statistics simpler if they separated descriptive statistics from statistical modelling and inference. Chapters 11 to 13 are purely on descriptive statistics. There we describe graphs in Chapter 11, data summary in Chapter 12 and tabulation in Chapter 13.

In Chapters 14 and 15 we introduce the basic ideas of statistical modelling. Chapter 14 describes some probability ideas, while Chapter 15 covers basic concepts of statistical inference. These Chapters are important in their own right, but they also serve to introduce they key ideas that will be used in the later chapters. We find that a stumbling block for many non-statisticians is that they are hazy about these ideas.
Chapter 11 - Graphs

11.1 Introduction

Graphs, sometimes called plots or charts, are for three purposes. Exploratory and diagnostic graphs are described in Sections 11.2 and 11.3 and are to help the person who is doing the analysis. Graphs can support the teaching of statistics as we explain in Section 11.4. Finally we have presentation graphs. These are as part of the report or seminar, and are designed to show the results to readers, or viewers.

Instat’s graphics menu is shown in Fig. 11.1a and we will mainly describe the options in this menu.

Graphs can also be generated as part of the output from many of the dialogues in the statistics menu and these are described in later chapters. As an example we show part of the Statistics ⇒ Regression ⇒ Simple dialogue in Fig. 11.1b. This dialogue is considered in Chapters 15 and 17 and includes many options for graphs, for all three purposes considered in this Chapter.

11.2 Exploration

Scatter-plots can be with points or lines, and use the Graphics ⇒ Plot dialogue. The first example is for the experimental data, described in Chapter 3, Section 3.4 which we show again in Fig. 11.2a. To help some of the later plots we have 2 extra columns in the Experiment worksheet, and Fig. 11.2a shows that they contain the mean of X1, for each of the 3 treatments. We used the Statistics ⇒ Data Summary ⇒ Column Statistics dialogue to give these means, and describe this dialogue in Chapter 12. For now, if you are following this analysis, you could just enter these data into x4 and x5 as shown in Fig. 11.2a. We also note, from Fig. 11.2b, that the overall mean of x1 is 332.75.
In **Fig. 11.2c** we show the simplest use of the Graphics ⇒ Plot dialogue. We plot X1 against X3 and the resulting graph (with a little poetic licence) is shown in **Fig. 11.2d**.

The data set also includes the four blocks. When there is structure to the data, it is useful to be able to look at this structure in the graph. We show the dialogue in **Fig. 11.2e**. The new feature is that we have included the blocks as a "By factor" in the dialogue. This column must have been defined as a factor previously. The results, (with a little more poetic licence) are shown in **Fig. 11.2f**.
We now explain the "poetic licence", namely how did we get (blue) circles in the first plot, Fig. 11.2d, and dotted lines in the second, Fig. 11.2f? In Fig. 11.2c and Fig. 11.2e there is a button labelled "Define lines and symbols". When this is pressed you get a submenu, as shown in Fig. 11.2g for the first graph and Fig. 11.2h for the second.

The way that this works is to send an instruction to Instat, to plot the y-variable with the specified symbol and line for all future plots. So, for any y-variable, you do not need to use this sub-dialogue each time, but only when you want to change the symbol or the line style.

Sometimes you may wish to look at the raw data together with summary statistics. The full plot dialogue is shown in Fig. 11.2i and we here have an overlay of the means as shown in Fig. 11.2j.
A similar graph to Fig. 11.2j can be found from one of the statistics menus. In Chapter 16 we will introduce the Analysis of Variance and part of one of the dialogues is shown in Fig. 11.2k. With the options set, this produces the graph shown in Fig. 11.2l.

Fig. 11.2k Stats ⇒ ANOVA ⇒ One Way Fig. 11.2l Results

Boxplots are also a valuable exploratory tool. In this version of Instat they are not in high-resolution, but are still useful. The example in Fig. 11.2m uses the data from the worksheet called rain, that has 32 years of monthly rainfall totals from a site in Sri Lanka.

Boxplots show the general shape of a set of data and indicate possible outliers, as shown in Fig. 11.2m.

11.3 Diagnostic plots

There is no clear distinction between data exploration and diagnostics. Roughly exploration come first and is to look at the data, before fitting a model, while diagnostic plots provide checks on the model that is being fitted.

Many analyses assume that the data come from a normal distribution. Hence one popular graph is called a probability plot. The dialogue in Fig. 11.3a is from Graphics ⇒ Probability Plots. We give three examples using the monthly rainfall worksheet described in Chapter 3. The data for August, in Fig. 11.3b look roughly normal, while those for November, in Fig. 11.3c do not. The annual totals look reasonably normal except for one extraordinary year, 1963.

**Fig. 11.2m Results from Graphics ⇒ Boxplot with rain.vor**

**Boxplots for Aug., Sept., Oct., Nov., Dec.**

```
X9  ------ ( + )--------
     ---------------------
X10  ------ ( + )------
     ---------------
X11  -----------------( + )---
     ---------------
X12  ------ ( + )--------
     ---------------   * *
X13  ------ ( + )------ o
     ---
```

Boxplots show the general shape of a set of data and indicate possible outliers, as shown in **Fig. 11.2m**.

**11.3 Diagnostic plots**

There is no clear distinction between data exploration and diagnostics. Roughly exploration come first and is to look at the data, before fitting a model, while diagnostic plots provide checks on the model that is being fitted.

Many analyses assume that the data come from a normal distribution. Hence one popular graph is called a probability plot. The dialogue in **Fig. 11.3a** is from Graphics ⇒ Probability Plots. We give three examples using the monthly rainfall worksheet described in Chapter 3. The data for August, in **Fig. 11.3b** look roughly normal, while those for November, in **Fig. 11.3c** do not. The annual totals look reasonably normal except for one extraordinary year, 1963.
Usually, however the assumption of a normal distribution applies to the residuals, after fitting a model, and not to the data themselves. Hence most of the diagnostic plots are options in the modelling dialogues. We consider them mainly in Chapters 16 and 17 of this guide.

In this section we have just shown one example of a diagnostic plot. Other plots are used to investigate different aspects of the model that is assumed, such as equal variance, and time dependence. These are considered in later chapters.

### 11.4 Teaching

In this Section we look at graphical features that are particularly adapted to support teaching.

The **Graphics ⇒ Plot** dialogue can be used to plot simple functions as well as data columns. We illustrate first with a non-statistical example.

We create or open a worksheet and put the two functions into Instat strings, S1 and S2, as shown in Fig. 11.4a. They are a quadratic and a linear function, namely

- \( X^2 - 6X + 4 \)
- \( 10X - 4 \)

Then we use the **Graphics ⇒ Plot** dialogue as shown in Fig. 11.4b. In this dialogue we have given the x-range as 0 to 20, as we have no x-variable. We have added the option to include some horizontal reference lines, to clarify the graph.
The resulting graph is in **Fig. 11.4c**.

**Fig. 11.4a** Entering formulae

**Fig. 11.4b** Graphics ⇒ Plot

The resulting graph is in **Fig. 11.4c**.

**Fig. 11.4c** Results for S1 and S2

**Fig. 11.4d** Results for S3 and S4

In **Fig. 11.4a** we also included the formula for an exponential distribution in S3 and a normal distribution in S4. We also included a constant, k1, in the formulae, which we set as 10.

The plot of S3 and S4 is shown in **Fig. 11.4d**, together with reference lines at the mean and at one standard deviation from the mean.

In case you wonder how the reference lines, in **Fig. 11.4d** "knew" to be at one standard deviation from the mean, we must first know that the standard deviation of the exponential distribution is equal to the mean. Then, in **Fig. 11.4e**, we show the output window, which has recorded the commands as we used the Manage ⇒ Calculate and the Graphics ⇒ Plot dialogues.

**Fig. 11.4e** Commands for graphs

```plaintext
k1=10
k2=k1-sqrt(k1)
k3=k1+sqrt(k1)
plot z3 z4;xax 0 20;vref k1-k3;nolegend
```
We rarely use Instat's commands in this Introductory Guide, but we indicate their role in Appendix 1. The example in Fig. 11.4e could be built into a small macro, that asked the user to specify the mean, and then gave the corresponding graph.

In Fig. 11.4a we showed four examples of formulae that are acceptable to Instat. The rules are quite strict and if Instat can not evaluate a function it will refuse to give the graph. The variable must be given by “x” or “X”. You may include standard functions, like “exp” or “sqr” as shown above, but can not include statistical functions, like mean(x3).

You can include “constants”, k1, k2, etc. in the formulae, but not columns such as x1.

There are more examples of Instat’s functions for teaching probability ideas in Chapter 14, where we introduce the Statistics ⇒ Probability Distributions dialogue. One example is shown in Fig. 11.4f and the resulting graph is in Fig. 11.4g.

This is a much easier way to graph the normal, or other distributions, than having to type the density functions yourself. If you are disappointed with some aspects of the graph, like the omission of reference lines, then Fig. 11.4h shows the functions that are in the worksheet now. We see that Instat has simply written the formulae for you, here into S5 and S6, and then plotted the functions, as we have described above. If you do not want these functions to be saved in the worksheet, then tick the “Remove after fit” option in Fig. 11.4f. But saving the formulae has the advantage that you could repeat the graph with the options that you would like.
We conclude this section with a different type of graph. In the next chapter we consider simple descriptive summaries. The stem and leaf plot is a useful graph and is similar to a histogram. We illustrate with the rice survey data (survey.wor) described in Chapter 3 and Fig. 11.4i shows the Graphics ⇒ Stem and Leaf dialogue.

**Fig. 11.4i** Graphics ⇒ Stem

The results are shown in Fig. 11.4j and 11.4k, both with, and without the histogram option. We find this display to be useful to explain that they are similar, and a histogram is like the stem plot without so much information in the “leaves”.

**Fig. 11.4j** Results –conventional display  
**Fig. 11.4k** Results with histogram option

In addition, the usual stem and leaf plot lists the outliers separately, while a standard histogram does not.

### 11.5 Presentation

In the introduction to this chapter we stressed the difference between exploratory or diagnostic plots on the one hand and presentation graphs on the other.

When a part of the analysis is completed then the report, or presentation is prepared, and is likely to contain a variety of graphs. As they are being prepared for others, it is useful to spend time editing the graph, so it is as clear as possible. Users can spend many happy hours changing the legend, or adding more tick marks, or changing the fonts for titles and so on.
One of our guides is called "Informative Presentation of Tables, Graphs and Statistics" and includes suggestions for graphical presentation of results.

If you use Instat you are saved all this time, because the current version does not have any facilities for editing graphs! It also does not have some common types of graph that are often used in presentations. We are not sad about the omission of pie charts, but would like to bow to the inevitable, and include bar charts in a future version of the software.

We find that Instat's graphs are acceptable for student projects and for draft reports. To print a graph, make it the active window and use File ⇒ Print, possibly with Print Preview first. To copy a graph to the clipboard, simply use Edit ⇒ Copy (or Cntl<C>). Then in a reporting package, such as Word, use Edit ⇒ Paste Special, NOT the simple Edit ⇒ Paste. Say you would like to import the picture.

An alternative is to use special software that can capture screens. We used Paintshop Pro for the figures in this guide.

When you need proper presentation graphs, we suggest that you export the data for the graph, from Instat into the usual package that you use. In the next section we mention a little about graphics facilities in other software.

11.6 In conclusion

In this chapter we have shown that Instat's graphics are useful for exploratory and for teaching purposes. Currently they do not have the capabilities or the flexibility for presentation graphics, of most spreadsheets or of the other statistics packages.

Many people use a spreadsheet for their graphics. Excel's graphics are very powerful and flexible. They do omit some standard graphs, such as boxplots, that are in all common statistics packages, but these can be provided by the many add-ins that supplement Excel's capabilities.

Excel also does not have an automatic facility for multiple graphs, if the data are in "list format", but this can again be provided through add-ins, such as SSC-Stat.

If you use Excel for presentation graphs, then we strongly suggest that you do not use the defaults. It is easy to add your own graphics templates so that your new defaults provide "good graphics". We have a guide titled "Guidelines for good statistical graphics in Excel"

This provides general guidance on the use of Excel for presentation graphs and is downloadable from www.rdg.ac.uk/~sns97aai.

The Excel graph in Fig. 11.6a shows the annual data on the flow of water in the river Nile. The graph is from Grubb and Robson (2000) and this example is considered in more detail below.

**Fig. 11.6a Annual river flow**

![Annual river flow graph](image-url)
Most standard statistics packages are strong enough in graphics that they can be used for both exploration and presentation. Fig. 11.6b and 11.6c were produced in Minitab and use boxplots to show the monthly river flow, and the residuals from a trend. One detail in these graphs, that is difficult to arrange with some software, is that the x-axis starts in May, i.e. the month numbers are not in numerical order. These graphs show the annual cycle as well as the variability of the data within each month.

**Fig. 11.6b** Boxplots of flow (log scale)  
**Fig. 11.6c** Boxplots of residuals

In Fig. 11.6d we show a simple Minitab graph of the residual values from one of the months, which shows that in June, a month with below average flow (i.e. negative residual values), these were becoming less negative through this record.

**Fig. 11.6d** Residuals for June

One feature in some packages that is not in Instat, nor Excel, is the facility to include multiple graphs on the same frame. Fig. 11.6e shows the 12 sets of monthly values, centred about their mean residual. In Fig. 11.6e the values for month 6 are the same as shown in the individual graph in Fig. 11.6d. We see, for example, that the trend in the residuals in June is not a feature of the data in the Autumn months. There is also an obvious outlier in November. This was noted as probably a data coding error, as the recorded value was 43.3, when 433 would have been a more reasonable value.
On the statistical side, these graphs, and the paper by Grubb and Robson (2000) illustrate the extent to which descriptive statistics can be taken. It is an example of a time-series analysis that is useful in its own right and also shows that standard time series models are unlikely to be very successful for this set of data. We look very briefly at time series modelling in Chapter 19, but the fact that the trend in the data is markedly different in the Summer and the Autumn is not a feature of standard time series models.

This is a good example, to show those who feel that statistics is now simply a question of finding a powerful statistics package and clicking on the different options. They are unlikely to find one that corresponds to the type of model needed here. The solution is not to give up modelling, but to give up the idea that data analysis can be reduced to a simple "clicking" exercise.

Finally, the statistics package S-Plus has promoted the idea of trellis-plots, and an example is in Fig. 11.6f. These are now available in some other packages and provide an excellent exploratory tool, particularly when the dataset is large. In such cases there is usually a considerable degree of structure, i.e. factor columns, and the trellis plots allow the data to be graphed in ways that includes this structure.
Fig. 11.6f  Example of a trellis plot

plot of $Y$ vs $X$ for each combination of block and treatment

Treat: A  
block: 1  
92  93  94  95  96  97

Treat: B  
block: 1  

Treat: C  
block: 1  

Treat: A  
block: 2  
92  93  94  95  96  97  

Treat: B  
block: 2  

Treat: C  
block: 2  

year  
measured response
Chapter 12 - Data summary

12.1 Introduction

Descriptive statistics, are invariably required during the initial phase of data analysis, usually together with graphical presentations. In some studies that is all that is needed for the analysis.

In this chapter we look at Instat's dialogues for data summary, for which the menu is shown in Fig. 12.1a. We have already seen some, particularly the dialogue for row statistics in Chapter 8 and for column statistics in Chapter 9.

12.2 Displaying descriptive statistics

The Statistics ⇒ Summary ⇒ Describe dialogue, shown in Fig. 12.2a, provides a simple way to display a range of summary statistics. In Fig. 12.2a we illustrate its use on one month of the rainfall data introduced in Chapter 3. Here we provide for the default display, to which we have added selected percentage points.

Some of the data and the results are shown in Fig. 12.2b. For example we see that the driest April had 18.7mm, while the mean for April was 217mm, slightly larger than the median, (the 50th percentile).

When further statistics are needed they may be requested, as is shown in Fig. 12.2c, together with the output in Fig. 12.2d.

28/06/2002
In Fig. 12.2e and 12.2f we show the formulae used for these summary statistics.

**Fig. 12.2e  Formulae for summary statistics - example data and notation**

<table>
<thead>
<tr>
<th>Data</th>
<th>Example</th>
<th>Ordered Data</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>10</td>
<td>$x_{(1)}$</td>
<td>3</td>
</tr>
<tr>
<td>$x_2$</td>
<td>15</td>
<td>$x_{(2)}$</td>
<td>10</td>
</tr>
<tr>
<td>..</td>
<td>14</td>
<td>..</td>
<td>11</td>
</tr>
<tr>
<td>..</td>
<td>12</td>
<td>..</td>
<td>12</td>
</tr>
<tr>
<td>..</td>
<td>17</td>
<td>..</td>
<td>14</td>
</tr>
<tr>
<td>..</td>
<td>3</td>
<td>..</td>
<td>15</td>
</tr>
<tr>
<td>$x_n$</td>
<td>11</td>
<td>$x_{(n)}$</td>
<td>17</td>
</tr>
</tbody>
</table>

**Fig. 12.3f  Formulae**

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>$n$</td>
<td>7</td>
</tr>
<tr>
<td>Minimum</td>
<td>$x_{(1)}$</td>
<td>3</td>
</tr>
<tr>
<td>Maximum</td>
<td>$x_{(n)}$</td>
<td>17</td>
</tr>
<tr>
<td>Range</td>
<td>$x_{(n)} - x_{(1)}$</td>
<td>14</td>
</tr>
<tr>
<td>Mean</td>
<td>$\bar{x} = \frac{\Sigma(x/n)}{n}$</td>
<td>11.71</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{(n - 1)}}$</td>
<td>4.54</td>
</tr>
<tr>
<td>Pth percentile</td>
<td>$P^* (n + 1)/100$th observation</td>
<td>4.54</td>
</tr>
<tr>
<td></td>
<td>$x_{(1)}$ if $P^<em>(n+1)/100 &lt; 1$, $x_{(n)}$ if $P^</em>(n+1)/100 &gt; n$</td>
<td></td>
</tr>
<tr>
<td>MEDian</td>
<td>$P_{(50)}$</td>
<td>12</td>
</tr>
<tr>
<td>LQUartile</td>
<td>$P_{(25)}$</td>
<td>10</td>
</tr>
<tr>
<td>UQUartile</td>
<td>$P_{(75)}$</td>
<td>15</td>
</tr>
<tr>
<td>IQUartile</td>
<td>$P_{(75)} - P_{(25)}$</td>
<td>5</td>
</tr>
<tr>
<td>SError</td>
<td>$s / \sqrt{n}$</td>
<td>1.71</td>
</tr>
<tr>
<td>SKEwness</td>
<td>$\sqrt{n} \frac{\Sigma(x - \bar{x})^2}{(\Sigma(x - \bar{x})^2)^{3/2}}$</td>
<td>-0.911</td>
</tr>
<tr>
<td>KURTosis</td>
<td>$(n \Sigma(x - \bar{x})^4) / (\Sigma(x - \bar{x})^2)^{3/2} - 3$</td>
<td>0.08</td>
</tr>
<tr>
<td>CSS (corrected ssq)</td>
<td>$\Sigma(x - \bar{x})^2$</td>
<td>123.4</td>
</tr>
<tr>
<td>USS (uncorrected ssq)</td>
<td>$\Sigma x^2$</td>
<td>1084.0</td>
</tr>
<tr>
<td>COEfficient of variation</td>
<td>$(s/\bar{x}) * 100%$</td>
<td>38.7%</td>
</tr>
</tbody>
</table>
12.3 Saving summary statistics

Summary statistics are often required, not merely as an end product, but to be used in further analyses or presentations. The **Statistics ⇒ Summary ⇒ Column Statistics** dialogue may be used to present specified summary statistics and optionally they may be saved back to the worksheet. An example, again for the 32 years of monthly rainfall, is shown in **Fig. 12.3a**, with the results in **Fig. 12.3b**. We see that the means for each month have been saved in x14, as well as being displayed in the output window.

![Fig. 12.3a Stats ⇒ Summary ⇒ Column Stats](image)

![Fig. 12.3b Results](image)

From here, adding a column with the numbers 1 to 12 for the months, it is simple using the ideas from the last chapter, to give the graph shown in **Fig. 12.3c**.

![Fig. 12.3c Graph of results](image)

In addition to the statistics shown in **Fig. 12.3a**, percentiles or proportions can also be given with this dialogue. For example we could ask for the proportion of years in which each month's total was less than 200mm.

Data are often structured into groups and then it is usually interesting to compare the different groups. We have already seen some of Instat's facilities for making graphical comparisons among groups in Chapter 11. The **Graphics ⇒ Plot** and **Boxplot** dialogues each have the facility to process data subdivided by the levels of a factor and the same is
available with the Column Statistics dialogue. Then the results consist of the requested summary statistics separately for each level of the classifying factor. As an example, Fig. 12.3d and Fig. 12.3e show the dialogue and results for the count and mean yields of rice in the survey data, subdivided by each variety.

These are examples of 'one-way' tables. Data, particularly from surveys are often summarised by both one-way and multi-way tables. The facilities for multi-way tables are described in the next chapter.

12.4 More one-way tables

Sometimes it is useful to look at the counts or percentages of a factor, or to group the values of a variate into a frequency distribution. For the survey data, we show the Statistics ⇒ Summary ⇒ Group dialogue in Fig. 12.4a, to look at the fertiliser amounts applied. The results are in Fig. 12.4b and show that 9 of the 36 farmers did not apply any fertiliser. Also 25% of the farmers applied more than 2cwt/acre of fertiliser.
The values for fertiliser are discrete, in that only integer, or half-integer values are used. In Fig. 12.4c and d we look similarly at the frequencies for the yields, but this time we give the “boundary values” shown in Fig. 12.4c.

In Fig. 12.4d we see that a quarter of the farmers had yields of 30 cwt/acre or less, and just 2 farmers had yields of more than double this value.

One reason for grouping data into a frequency distribution is to compare the observed frequencies with the expected values from an assumed probability distribution. This dialogue gives the observed frequencies, while the Statistics ⇒ Probability Distributions dialogue, described in Chapter 15, is used to give the expected frequencies for a range of distributions. Then the Statistics ⇒ Simple Models ⇒ Goodness of Fit dialogue is used to give the chi-square test to compare them.

12.5 In conclusion

All statistics packages offer the same type of summary statistics options as those shown in Sections 12.2 and 12.3. With Minitab, the menu, shown in Fig. 12.5a, distinguishes between the display and storage of the summaries in the same way as Instat.

The equivalent menus for Excel, using SSC-stat is shown in Fig. 12.5b.
Excel is different, because it does not distinguish between the spreadsheet and the output window. So the summary statistics are automatically in cells that can be used for further analysis if necessary.

The other distinction between Sections 12.2 and 12.3 is the flexibility of defining which statistics to display. Using Excel with SSC-Stat the dialogue to specify the statistics to be evaluated is shown in Fig. 12.5c, and the results from the survey data are in Fig. 12.5d.

**Fig. 12.5 Summary dialogue in Excel**

**Fig. 12.5d Results**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Variety</td>
<td>Count</td>
<td>Mean</td>
<td>Stdev</td>
</tr>
<tr>
<td>2</td>
<td>NEW</td>
<td>4</td>
<td>59.6</td>
<td>2.55</td>
</tr>
<tr>
<td>3</td>
<td>OLD</td>
<td>17</td>
<td>45.4</td>
<td>7.13</td>
</tr>
<tr>
<td>4</td>
<td>TRAD</td>
<td>16</td>
<td>30.0</td>
<td>6.52</td>
</tr>
</tbody>
</table>
Chapter 13 - Tables

13.1 Introduction

The production of summary tables is an important component of the analysis of many sets of data, particularly those from surveys. The "good practice" guide titled "Approaches to the Analysis of Survey Data" describes the way in which surveys are usually analysed, while the guide titled "Modern Approaches to the Analysis of Experimental Data" shows that the production of tables is also an important component of the analysis of experimental data. The route to Instat's tabulation facilities is shown in Fig. 13.1a.

We mainly use the small survey of rice yields, introduced in Chapter 3, to illustrate Instat’s tabulation facilities. However tables are general and we therefore also include examples from experiments and from monitoring data.

Tables, like graphs, can be used either as exploratory tools or to present results. The presentation of the results is shown in most sections of this chapter, and the use of tables to look critically at the data is described in Section 13.3.

A further use of tables is to provide an initial summary of the data that can then be used for further analysis. This is described in Section 13.11. We have already seen this type of summary with the Statistics ⇒ Data Summary ⇒ Column Statistics dialogue. The facilities here extend this idea, slightly in the range of summary statistics and particularly because the summary can be over more than one factor.

In Section 13.2 we describe how frequency tables can be produced, that contain either counts or percentages. We produce tables with summary statistics in Section 13.3. There we also show one way of data exploration, by introducing a "drill-down" facility that shows the summaries together with the raw data.

In Section 13.4 we describe ways in which tables can be manipulated, and formatted so the results are presented as clearly as possible. Here the parallel with the use of graphs is more obvious, because many users are familiar with the ways in which graphs can be tidied to present the results as informatively as possible. As a presentation tool we also show how the table can be made to resemble a bar chart, to help in the interpretation of the data.

We believe that training courses could introduce and review many statistical ideas through the use of tables. This is partly because large and hence potentially interesting datasets can be investigated and helps the training course to devote more time to teaching descriptive statistics. In Section 13.5 we describe teaching ideas, particularly for teaching percentages.

It is often useful to weight the information when forming a table, because the observations represent different proportions of the population. In Section 13.6 we introduce the idea of weighted tables.
We show the General Tables dialogue in Section 13.7. This combines and extends the facilities for producing tables that are described in Sections 13.2 and 13.3. With this dialogue we can produce tables for as many summary statistics as we like and have a wide range of ways that the data can be summarized.

Surveys sometimes include questions for which there are multiple answers, for example “Name up to 3 health problems you have had in the last 6 months.” In Section 13.8 we show how multiple response data can be coded and processed. In Section 13.9 we look more generally at processing data when information is at multiple levels.

### 13.2 Frequency tables

The layout of a table is determined by the factors that make up the rows and columns. These factors must first be defined, using the Manage ⇒ Data ⇒ Make Factor dialogue. This step is done only once, because Instat keeps the information internally on which columns are factors.

The rice survey data was introduced in Chapter 3 of this guide and the data are in the Instat worksheet called survey.wor. The 7 columns of data are as shown in Fig. 13.2a.

**Fig. 13.2a** Columns of data in survey worksheet

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Village</td>
<td>Factor with 4 levels (Sabey, Kesen, Niko and Nanda)</td>
</tr>
<tr>
<td>X2</td>
<td>Field</td>
<td>Field number</td>
</tr>
<tr>
<td>X3</td>
<td>Size</td>
<td>From 1 to 30 acres</td>
</tr>
<tr>
<td>X4</td>
<td>Fert</td>
<td>Quantity, from 0 to 3 cwts per acre</td>
</tr>
<tr>
<td>X5</td>
<td>Variety</td>
<td>Factor with 3 levels (New, Old and Trad)</td>
</tr>
<tr>
<td>X6</td>
<td>Yield</td>
<td>Ranging from 19.1 cwts/acre to 62.1 cwts/acre</td>
</tr>
<tr>
<td>X7</td>
<td>Fertgrp</td>
<td>Fertiliser group, labelled 0cwt, 0.5 to 2cwt and &gt; 2 cwt</td>
</tr>
</tbody>
</table>

The Village (X1) and Variety (X5) columns were designated as factors and we will mainly use these columns to form the tables. The simplest, shown using the Statistics ⇒ Tables ⇒ Frequency dialogue, is a table of counts, or frequencies, shown in Fig. 13.2b with the results in Fig. 13.2c.

**Fig. 13.2b** Statistics ⇒ Tables ⇒ Frequency

**Fig. 13.2c** Counts

Tables can also be used to display percentages. There are 3 options, all of which are shown, together with the counts, in Fig. 13.2d.
As an example to explain the different percentages, consider the 3 farmers growing traditional rice in the village Sabey. This represents 30% of the farmers in Sabey (top right), and 20% of the farmers growing the traditional variety (bottom left). Finally it represents 8% of the 36 farmers in the study.
13.3 Summary tables

Tables can also provide summaries of other columns. This uses the Summary Tables dialogue, shown in Fig. 13.3a. The results, in Fig. 13.3b give the mean yield of the farmers, tabulated by the village and variety.

Fig. 13.3a  Statistics ⇒ Tables ⇒ Summary

Fig. 13.3b  Mean yields

Here we see that the overall mean yield was 40.6 cwts/acre. We also see that the mean yields for each variety are different, ranging from 59.6 for the new improved, to 30 for the traditional rice. We also find that there are considerable differences between the means in each village. However, the apparent village differences may simply reflect the fact that only 2 of the villages, Sabey and Nanda, had farmers who grew the best yielding variety.

In Fig. 13.3b the margins of the table are in gray. For example 45.3 is the mean yield for the 10 farmers in the village, Sabey. It is not therefore the mean of the three values in the body of the table, because they are each from different numbers of farmers. We know how many farmers, from the frequency tables in Section 13.2. Combining both tables the mean yield for the Sabey farmers, given in Fig. 13.3b, is:

\[(2 \times 59.4 + 5 \times 49.2 + 3 \times 29.5)/10 = 45.3\]

It is possible for tables to have more than two “dimensions”, that is to use more than two factors. Here there is not really enough data, but for illustration Fig. 13.3c shows the dialogue for the same table subdivided further by the quantity of fertilizer applied.

The results, in Fig. 13.3d, have a sort of triangular shape, that indicates that the farmers using the newer varieties also use more fertilizer.
It is also possible to display more than one summary in a table, as is shown in the dialogue in *Fig. 13.3e*. The results, in *Fig. 13.3f*, show this to be an alternative way, compared to *Fig. 13.3d*, of looking at the fertilizer and yields in the same table. The fertilizer can either be coded as a factor, *Fig. 13.3d*, or tabulated as a variate, *Fig. 13.3f*.

As a general concept, a comparison between *Fig. 13.3f* and *13.3d* shows that presenting multiple summaries in a 2-way table is like a 3-way table. This is part of the philosophy in Instat, but there is one difference in the two tables. Multiple statistics do not (usually) have a margin as can be seen above. There is a grey column margin in *Fig. 13.3d*, corresponding to the "Fertgrp" factor. But there is no column margin in *Fig. 13.3f*, since it is meaningless the combine the values of the three summaries (count, fertilizer use and yield).

A further useful feature of the summary tables dialogue is shown in *Fig. 13.3g*. This provides a sort of "drill-down" as shown in *Fig. 13.3h*. Here we are looking particularly at the
3 farmers in Sabey, who are growing the Traditional rice. As well as the summaries, we see here the individual observations that make up this summary. This is a useful exploratory tool to look for anomalies in the data, or to understand more of the pattern in the data. Here we see that the mean yield of 29.5 is made up from the 3 observations 33.6, 30.6 and 24.3. We also see that the lower yield of 24.3 corresponds to a farmer who did not apply fertilizer.

**Fig. 13.3g Drill down**

**Fig. 13.3h Results**

Usually tables show the summary of the data, but they can sometimes show all the individual values. In Chapter 16, one of the more complicated examples, in a worksheet called “beans.wor”, has four treatment factors plus the blocks. The dialogue is given in **Fig. 13.3i** and the five-way table, that shows the data is in **Fig. 13.3j**. It provides a clear way of looking at the values together with all the treatment and the block means. The summary, in **Fig. 13.3k**, is derived from this table, using the methods to be shown in Section 13.4. It presents the means for the three factors that were shown to be important in the analysis of variance given in Chapter 16.

**Fig. 13.3i Five-way table**

**Fig. 13.3j Results**
13.4 Presenting tables effectively

So far we have shown many tables that arise directly from the Instat tables dialogues. In this section we show how the information in the resulting table can be manipulated and formatted. This is, in many ways, the parallel of being able to edit a graph so that the message is displayed as clearly as possible.

The two key menus when a table is the active window are shown in Fig. 13.4a.
Fig. 13.4a Structure and format menus

Fig. 13.4b Table after formatting

<table>
<thead>
<tr>
<th>Village</th>
<th>Variety</th>
<th>Mean Feet</th>
<th>Mean Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td></td>
<td>2.3</td>
<td>59.4</td>
</tr>
<tr>
<td>SABEY</td>
<td>CLD</td>
<td>2.3</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>0.8</td>
<td>29.5</td>
</tr>
<tr>
<td>NANDA</td>
<td>CLD</td>
<td>1.9</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>1.7</td>
<td>36.3</td>
</tr>
<tr>
<td>KostNiko</td>
<td>CLD</td>
<td>1.3</td>
<td>40.4</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>0.1</td>
<td>25.7</td>
</tr>
</tbody>
</table>

Fig. 13.4b has been produced, using these dialogues, starting from the table shown in Fig. 13.3f, and shows some of the many ways we can manipulate information in the table.

From the Structure ⇒ Levels dialogue, shown in Fig. 13.4c, we have omitted one “level” of the summary, namely the counts. We have also changed the order of the Villages in the display.

Then we used Structure ⇒ Merge, Fig. 13.4d, to combine the information from 2 of the villages and Structure ⇒ Margins, Fig. 13.4e, where we have chosen not to display the margins. At this stage, our table looks as shown in Fig. 13.4f.
From the **Format ⇒ Cells** dialogue, shown in **Fig. 13.4g**, we display the information in a way that resembles bar charts. Finally from **Format ⇒ Grid**, see **Fig. 13.4h**, we have changed the size of the cells in the grid, so the results are shown more clearly. This gives the table, shown above, in **Fig. 13.4b**.

Results can also be scaled, or the origin changed. For example, the means in the table above are in the old-fashioned units of cwt per acre. Changing the scale could present them in tons/ha. Or, knowing the overall mean is 40.6, we could change the origin for the yields to 40.6, and the table would then show the average deviations from the overall mean.

In making these changes we have not lost the original information. So we can easily try different presentations to emphasize different aspects of the results.

The **Structure ⇒ Interactive** dialogue is shown in **Fig. 13.4i**. This is partly for teaching purposes and provides a quick way for users to see the types of changes that can be done on a table from a single dialogue.

From this dialogue the levels within a factor or factors themselves can be hidden, to emphasize those that remain. Similarly the table margins can be displayed or not. Parts of the table can be coloured and the user can set a priority level to indicate which settings have priority, if cells are coloured in more than one way.
This dialogue does not have the full flexibility of the individual dialogues that were used above, to produce Fig. 13.4b. That would have made it too complicated. For example “flooding”, which is a way of producing simple bar charts in the table, can only be done for percentage data, i.e. where the whole cell is coloured if the value is 100%. For other possibilities the Format ⇒ Cells dialogue is used.

Once you have a table, the defining factors should not be changed. But the data in a statistics column, can be changed, for example if mistakes were found, or if you want to see the effect of changing odd values into missing data. Similarly the contents of the specified weight or filter columns can be changed. Then use the Data ⇒ Rebuild Table menu to regenerate the table with the new data. Other table dialogues allow you to change the columns that are used for weighting or the filter column, once you have the table.

13.5 Teaching percentages and other topics

One of our aims in producing this module on tables is to support the teaching of statistics. We consider the teaching of percentages in this section and introduce the idea of data at multiple levels in Section 13.8 and 13.9, when multiple responses are discussed.

The general idea that a wide range of descriptive statistics can be produced and then studied. This helps courses to spend longer on the teaching of descriptive statistics and that is useful in its own right, because many analysis need little more than this. Extra time, devoted to descriptive statistics, will also help to give users a solid foundation from which to start the more demanding topics concerned with statistical modelling.

It is important for users to practice using tables and graphs imaginatively with data sets that are of a realistic size. We hope that the tabulation facilities in Instat, like the use of pivot tables in spreadsheets, such as Excel, can make it rewarding to produce and present data in tables.

One use of these tables is to teach the range of ways that percentages are used. They are commonly employed and sometimes not clearly understood.

The key to using percentages is to understand what constitutes 100%.

We have already shown the routine use of percentages, with the frequency tables, produced in Section 13.2. We repeat 2 of these tables here in Fig. 13.5a and 13.5b. In the first, the
number of farmers in each village provides the 100% and in the second, it is the number growing each variety.

**Fig. 13.5a Row percentages**

<table>
<thead>
<tr>
<th>Village</th>
<th>NEW</th>
<th>OLD</th>
<th>TRAD</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>SABEI</td>
<td>20</td>
<td>50</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>KESEN</td>
<td>0</td>
<td>43</td>
<td>57</td>
<td>100</td>
</tr>
<tr>
<td>NIKO</td>
<td>0</td>
<td>40</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>NANDA</td>
<td>14</td>
<td>50</td>
<td>36</td>
<td>100</td>
</tr>
<tr>
<td>All</td>
<td>11</td>
<td>47</td>
<td>42</td>
<td>100</td>
</tr>
</tbody>
</table>

**Fig. 13.5b Column percentages**

<table>
<thead>
<tr>
<th>Variety</th>
<th>NEW</th>
<th>OLD</th>
<th>TRAD</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>SABEI</td>
<td>50</td>
<td>23</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>KESEN</td>
<td>0</td>
<td>18</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>NIKO</td>
<td>0</td>
<td>12</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>NANDA</td>
<td>50</td>
<td>41</td>
<td>33</td>
<td>39</td>
</tr>
<tr>
<td>All</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

The summary tables dialogue, described in Section 13.3, can be used, as shown in **Fig. 13.5c**, to present some percentage points of the yields. They are sometimes called percentiles and are given, in **Fig. 13.5d**, for each variety and for the farmers overall. For example we see that the lower quartile, i.e., 25% of the farmers have a yield of less than 29.6 bu/acre. We have used the ideas described in Section 13.4 to edit the table, so the information is shown clearly.

**Fig. 13.5c Percentage points of the yield**

<table>
<thead>
<tr>
<th>Summary Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
</tr>
<tr>
<td>OLD</td>
</tr>
<tr>
<td>TRAD</td>
</tr>
<tr>
<td>All</td>
</tr>
</tbody>
</table>

**Fig. 13.5d Results**

Fig. 13.5d gives similar information to what is often shown graphically in boxplots, and it is sometimes useful to show both. Two points about the tables are that they include the actual numbers, and also they provide the margin. Here the margin gives the summary statistics for all farmers and this is in addition to the breakdown by variety.

Sometimes we might look at results relative to a fixed standard. An example is shown in **Fig. 13.5f** where we have taken 50 cwt/acre as the standard we would like to aim for as the mean yield. We then scaled the results (Tables ⇒ Data ⇒ Scale), see **Fig. 13.5e**. If 50 is to be taken as 100% then we multiply the yields by 2. The results are shown in Fig. 13.5g.

For example, we see that the median farmer overall, with a yield from Fig. 13.5d of 40.7, is 81% of our standard. All the farmers growing the new variety exceed the standard, as do the upper quartile of farmers growing the old variety.
Sometimes we want the percentage of a total. An example is shown in Fig. 13.5h, using rainfall data from Nigeria. The raw file, calls samsmall, contains the daily data. The dialogue shown in Fig. 13.5g has been used to get the monthly totals for 2 years, together with their percentage of the annual total. From Fig. 13.5h we see, for example, that May rainfall is about 12% of the annual total.

To clarify the results we have then merged the months to give the same results for each quarter of the year in Fig. 13.5i. This shows that well over half the rainfall is in the Summer months.
We accept that these are all simple ideas, and already taught. But we believe that it is important for users to practice presenting and interpreting results in tables, in the same way that is already familiar in graphs. People are used to spending time to “beautify” a graph. If the same time were devoted to making tables as informative as possible, this would support good data interpretation.

It is sometimes useful to show the actual values that are “behind” the table. In Section 13.3 we showed one “drill-down” facility and we show another in Fig. 13.5j. This uses the Data ⇒ Show Values or Sorted Values options, once the tables are displayed. The tool-tip can then be used to explore the data within the table. We illustrate with 3 examples below. When the table is of counts, we indicate which rows of data, or “cases” correspond to the summary. When there is a percentage, the tool-tip simply reminds us of the count.

Fig 13.5j (Table) Data ⇒ Show sorted values

The final example is particularly appropriate for teaching. The table on the bottom right of Fig. 13.5j is of the median yields. We have used the menu option to display the values sorted into ascending order and we see that 50.7 is the 3rd of the 5 observations. This can also be used for tables giving other percentage points, to reinforce their meaning.

13.6 Weights in tables

The size of the fields was also recorded in the survey considered in this chapter. This varies from 1 to 30 acres and gives us a choice in the way we calculate the mean yields. If our unit of interest is the farmer, then each observation is of equal value and we need an ordinary mean. But if our interest is the land, then the yield from a farmer with 30 acres is 30 times the importance of the yield of a farmer with just 1 acre.
We use the drill-down facility in Fig. 13.6a to show what we mean. There the mean of 29.5 for Traditional farmers in Sabey is given by

\[(33.6 + 30.6 + 24.3)/3 = 29.5\]

If the size of the field is taken into account then the mean yield is calculated as

\[(33.6 \times 1.5 + 30.6 \times 4.0 + 24.3 \times 3.5) = 28.6\]

Weights can be applied within the Statistics → Tables → Summary dialogue. Alternatively, the system of weighting can also be changed once the table has been calculated. This dialogue is shown in Fig. 13.6b and the full table of means, weighted by the size, is shown in Fig. 13.6c.

13.7 General tables

The General Tables dialogue combines and extends the options for producing tables compared to the Frequency Tables and Summary Tables dialogues that were described in Sections 13.2 and 13.3.

The dialogue can be used for all the aspects described so far and embodies many of the ideas of constructing tables that we would like to encourage with Instat. Users can add individual statistics as they wish. They may be counts, percentages, or other summaries, and each column could be weighted differently.

The example in Fig. 13.7a and 13.7b combines some of the ideas described in the earlier sections. The table gives the counts of farmers and also the counts weighted by the farm size. This is the number of acres (rather than the number of farmers) in the sample and could alternatively have been given as a simple summary of the total of the size column.
Then we have the un-weighted and weighted mean yields as discussed in the previous section.

**Fig. 13.7a Statistics ⇒ Table ⇒ General**

Other summary statistics can be provided and we look particularly at some that extend the discussion of percentages that was in Section 13.5.

**Fig. 13.7c Percentiles**  
**Fig. 13.7d Proportions**  
**Fig. 13.7e Results**

The example in **Fig. 13.7e** is for a one-way table of yields for each variety. Using the part of the dialogue shown in **Fig. 13.7c**, we first give the 80\(^{th}\) percentile. The results show that
overall 80% of farmers had yields of less than 52.4 bu/acre. Then, from Fig. 13.7d we give the proportion, expressed in percentages, of farmers with yields of less than 40 bu/acre. There we see that almost all traditional farmers (93%) had a yield of less than 40, while it was only 11% for the farmers growing old improved rice, and none for those growing the NEW variety.

13.8 Multiple responses

Multiple responses are where a set of columns give the responses to related questions. The simplest is called multiple dichotomy, for example

- Indicate which of the following activities, “walking, jogging, cycling”, you have done in the past week. We show the layout of the data in Fig. 13.8a.

Fig. 13.8a Data to illustrate multiple dichotomy

An alternative way of posing roughly the same question is to ask:

- From the list below, put in order of your use up to 3 activities you did in the past week. This is called a multiple response. The same data as above are shown in Fig. 13.8b.

Fig. 13.8b Data to illustrate multiple response

The third possible layout is to recognize that you have information at two levels. For this example these are the person and the activity level, as shown in Fig 13.8c.
**Fig. 13.8c Multiple response data, shows the two levels**

<table>
<thead>
<tr>
<th>Person level</th>
<th>Activity level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Sex</td>
</tr>
<tr>
<td>1</td>
<td>Male</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
</tr>
<tr>
<td>6</td>
<td>Female</td>
</tr>
<tr>
<td>7</td>
<td>Female</td>
</tr>
<tr>
<td>8</td>
<td>Female</td>
</tr>
<tr>
<td>9</td>
<td>Female</td>
</tr>
<tr>
<td>10</td>
<td>Female</td>
</tr>
<tr>
<td>11</td>
<td>Female</td>
</tr>
</tbody>
</table>

This third layout is the obvious way to organize the data if there is also supplementary information about each activity, for example how often it was done. We consider this layout of the data in the next section.

Multiple response data sometimes arise from direct questions, as in the example considered here. They also arise from open-ended questions. For example, the question above could be open-ended as in:

- Please write down (or tell me) what exercise you have taken in the past week.

In such cases the information has to be coded before analysis, and sometimes the answers correspond to more than one code.

Instat permits any of these three layouts of the data to be used as shown in the dialogue in **Fig. 13.8d**.

**Fig. 13.8d Statistics ➔ Tables ➔ Multiple Response**

The “Unique responses” checkbox above concerns the action to be taken if a person (often called a respondent) gives the same answer on multiple occasions. We already know that the same person can give multiple responses, i.e. they can both walk and jog. But what if
someone put “Walking” twice? Should this be counted as 2 responses (i.e. 2 votes for walking from the same person) or just one. Use the Unique checkbox to count it just once.

An example where responses might be counted more than once would be “Films you have seen in the last month.” Then there are multiple responses from anyone who has seen more than one film, and non-unique responses from anyone who has seen the same film twice.

The reason for this special dialogue is the need to tabulate the set of multiple responses against any of the ordinary factors. So a table might look as shown in Fig. 13.8e. We see that of the 11 responses, 8 were male and 3 were female. We see that 2 males cycled, 2 walked and 4 cycled.

Fig. 13.8e Count of responses

<table>
<thead>
<tr>
<th>Summary</th>
<th>Male</th>
<th>Female</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Jogging</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Cycling</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>All</td>
<td>8</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

Fig. 13.8f Counts showing respondents

<table>
<thead>
<tr>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Jogging</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Cycling</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Responding</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Respondents</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>All</td>
<td>11</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

This table gives a summary of the responses. But these data are at 2 “levels”. There were 8 respondents and 11 responses. This is shown more clearly in Fig. 13.8f, where we have added information about the respondents. We see that there were 5 males and 3 females who gave these 11 responses. The other extra line informs us that one male and one female did not take any exercise. So there were 6 people who gave the 11 responses. The extra rows of data in Fig. 13.8f resulted from ticking the “All respondents” and “All +ve respondents” boxes in Fig. 13.8d.

Percentages can also be given. Then you must define what constitutes 100%, and here you have the same 3 choices that we have just described above. They are shown in the dialogue in Fig. 13.8g, within the frame “100% represents all”.

Fig. 13.8g Giving percentages
The first option is to give the percentage of responses, as shown in Fig. 13.8h. This is conceptually simplest, because the tables of counts, shown above, are of responses. Hence this table is just at the response level.

**Fig. 13.8h % responses**

<table>
<thead>
<tr>
<th>Summary</th>
<th>Male</th>
<th>Female</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>25</td>
<td>67</td>
<td>36</td>
</tr>
<tr>
<td>Jogging</td>
<td>25</td>
<td>33</td>
<td>38</td>
</tr>
<tr>
<td>Cycling</td>
<td>50</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>Responses</td>
<td>8</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>All</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Alternatively you can take account of the 2-level nature of the data and consider the percentage of respondents, as shown in Fig. 13.8i. In this case the percentages do not necessarily add to 100%, because there may be multiple responses per respondent. When using the number of respondents, you may use all respondents; there are 8 in this example, or the number that gave at least one response. This third possibility is shown in Fig. 13.8j.

Using the data in the first cell in Fig. 13.8h, 13.8i, and 13.8j, shows:

- 25% (i.e. 2 out of 8) of the male **responses** were for walking
- 40% (i.e. 2 out of the 5 males) of the male **respondents** walk
- 50% (i.e. 2 out of the 4 males) walk, of those males who take exercise.

In this third case, the column total of 200% under Male, indicates that each male who takes exercise, does so, on average in 2 different ways.

These three alternatives apply either when giving the overall percentage, or when, as above, giving the percentage of one of the ordinary factors.

When giving the percentage of the multiple response, as shown in Fig. 13.8k, there is usually no difference between the results using responses or respondents as the base. They are only different if a respondent may give the same activity (e.g. walking) more than once, and the "Unique" checkbox is not ticked.

**Fig. 13.8k Percentages of multiple response column**

<table>
<thead>
<tr>
<th>Summary</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>Male</td>
</tr>
<tr>
<td>Jogging</td>
<td>67</td>
</tr>
<tr>
<td>Cycling</td>
<td>100</td>
</tr>
<tr>
<td>All</td>
<td>73</td>
</tr>
</tbody>
</table>

This section has also introduced yet more ideas of percentages, to add to those mentioned in Section 13.5. When combined over multiple levels they no longer need to add to 100%. But the key remains the same. In each case it must be clear what constitutes 100%.

### 13.9 Data at multiple levels

Although multiple response-type questions are common in surveys, the subject is rarely taught. With the exception of SPSS, that has extensive facilities for tabulating multiple responses, they seem hardly evident in most statistics packages.

The idea of multiple responses becomes simple if the two-level nature of the data is recognized. This is then a useful teaching tool, because many surveys and other studies...
collect information at multiple levels. For example an educational study may collect information at “classroom level” from the teachers, and at “person level” from the pupils.

We therefore show the example from Section 13.8 again, in Fig. 13.9a, where we have added more information to clarify the 2-level nature of the data. In addition to the age group of each respondent, we have information about the time for each activity. Usually, when data are at more than one level, then each level is held separately, perhaps on separate sheets in a spreadsheet, or in linked tables in a database package. In Fig. 13.9a, X4 is the linking column, and indicates which person did the activity.

**Fig. 13.9a Data at two levels**

<table>
<thead>
<tr>
<th>X1</th>
<th>X2 - F</th>
<th>X3 - F</th>
<th>X4</th>
<th>X5 - F</th>
<th>X6</th>
<th>X7 - F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Sex</td>
<td>Age</td>
<td>Per</td>
<td>Imp</td>
<td>Time</td>
<td>Sex2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Male</td>
<td>51</td>
<td>Walking</td>
<td>30</td>
<td>Male</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Male</td>
<td>&lt;21</td>
<td>Jogging</td>
<td>20</td>
<td>Male</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Male</td>
<td>&lt;21</td>
<td>Cycling</td>
<td>60</td>
<td>Male</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Male</td>
<td>21-50</td>
<td>Cycling</td>
<td>20</td>
<td>Male</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>Male</td>
<td>&lt;21</td>
<td>Jogging</td>
<td>15</td>
<td>Male</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>Female</td>
<td>51</td>
<td>Cycling</td>
<td>120</td>
<td>Male</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>Female</td>
<td>21-50</td>
<td>Walking</td>
<td>10</td>
<td>Male</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>Female</td>
<td>21-50</td>
<td>Cycling</td>
<td>20</td>
<td>Male</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>Walking</td>
<td>40</td>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>Jogging</td>
<td>30</td>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>Walking</td>
<td>20</td>
<td>Female</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Fig. 13.9b we show the multiple response dialogue again to produce an example of the same table that we saw in Section 13.8. This table of counts is in Fig. 13.9c.

**Fig. 13.9b Statistics ⇒ Tables ⇒ Multiple Response**

**Fig. 13.9c Results**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Jogging</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Cycling</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Responding</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Respondents</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>All</td>
<td>8</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

With the data layout in Fig 13.9a we can also use the ordinary tabulation dialogues that are in Excel, or in any statistics package. We can get tables at the person level – see columns X1-X3 above. Or we can get information at the activity level using the data from columns X4-X6.

We illustrate how we can combine data across levels, by copying the sex of the person, from X2 to enable the activity-level tables to be produced without any special facilities. We do not of course need to do this so much in Instat, because of the special multiple response dialogue, but it is useful to show, because it illustrates how any package can produce most of the tables in this section. Copying data between levels is easily done in most packages.
In Instat use the Manage ⇒ Transformations ⇒ Expand dialogue as shown in Fig. 13.9d. This was used earlier in Chapter 9, Section 9.6. The resulting column is shown in Fig. 13.9e.

Fig. 13.9d Manage ⇒ Transformation ⇒ Expand

Now we can use the ordinary frequency tables dialogue, described in Section 13.2 to get any tables at the response level. An example of the completed dialogue is shown in Fig. 13.9f and gives the table shown in Fig. 13.9g, the same as from the multiple response dialogue. Alternatively we could opt, from this same dialogue, to get percentages.

Fig. 13.9f Statistics ⇒ Tables ⇒ Frequency

Similarly, to analyse data columns of responses, at the respondent (or person) level, the first step is to summarise the relevant columns, perhaps to give their means per respondent. An example, from the illustration above, would be to summarise the time per person, spent on all their activities. In Instat this step could use the Statistics ⇒ Summary ⇒ Column Statistics dialogue. Then the Summary Tables dialogue, described in Section 13.3, could be used.

The only tables that are more awkward to get are those that combine the information at both levels in the same table. This concept is general. Whenever we have data at multiple levels and wish to combine the information over different levels there are likely to be problems.
We can simplify multiple level problems. We do this by bringing all the information to a single level, as we did with X7 above.

Alternatively, the multiple-response dialogue, used in this section and in Section 13.8, is one example of how information can be presented over two levels. In general the subject of “multi-level” analysis is a very rich but complicated subject. For teaching purposes it is good to have such a simple example so we can introduce the general topic of handling data at multiple levels, even while teaching descriptive statistics.

13.10 Missing values

One reason for using a statistics package, rather than a spreadsheet, for data analysis is that all statistics packages cope with missing values.

In the tables in Instat there can be missing values in any of 4 ways:

- There can be missing values in the factor columns that define the table. Then the row of data is omitted. There is no other possible action, because it is impossible to know to which cell of the table the data should be allocated.
- There can be missing values in the variates that are used to give the summary observations. The observation is then omitted from the calculation, so effectively the row is omitted. However, the calculation proceeds for that cell of the table if there are still non-missing values. So, if three out of the eight farmers who grow a variety give missing values, the mean yield is still calculated, based on the other five.
- There can be missing values in a column of weights. The weight is then set to zero, which omits that observation from the calculation. Where the weight is applied to the whole table this corresponds to omitting the whole row of data. Where the weight is applied to a particular summary, the observation is just omitted from that summary.
- There can be missing values in the filter column. A filter column is applied to the table as a whole, and a missing value is treated like zero, i.e. FALSE. So the row is omitted from the data.

Sometimes there are multiple summary statistics in the same table. Then the user has an option when there are missing values in any column that is to provide a summary statistic. The first option is to remove the row of data completely from the calculation. Then the user knows that all summaries in a given table are based on the same rows of data, for example on the same people. The alternative is to omit the row for the calculation of that particular summary.

Instat has 3 missing value codes, though most examples use only one. When more than one is used, the two options above can be applied differently to each code.

Where there are missing values in the factors, weights or the filter column, this is included in the report of the table that is normally written to the output or log window. For summary columns one of the counts in the General dialogue is of the missing values, so users can see how much is missing in each column.

13.11 Saving the information in the table

For illustration we use the simple table from the survey data, shown in Fig. 13.11a.
Within the table menus we use File ⇒ Save, or Save As. The different ways we can save a table are then shown in Fig. 13.11b and we illustrate first with the html format. This is designed for the web, but it is also a format that can be imported into Excel, with much of the formatting intact, as is shown in Fig. 13.11c.

If we save instead as a comma separated file, then this can easily be imported anywhere, but comes without the formatting and labels, as is shown in Fig. 13.11d. This is also what results from using Edit ⇒ Select All and then Edit ⇒ Copy, within Instat, and then pasting into another package.

Either htm or rtf files can be imported into a word processor, as is shown from the rtf file in Fig. 13.11e. In this guide it looks similar to the previous figures, but the key difference is that it is a Word table, rather than a picture, and hence can be edited within Word in the normal way.

Most of the table formatting is kept in the rtf or html formats. One exception is flooding, which we have used in earlier sections to give a bar-chart type of presentation.

You can also save the table in what we have called Instat's table format, which then has the extension, "itb". This can then be loaded back into Instat on a later occasion and would look identical to Fig. 13.11a. However, it is no longer connected to the data in the Instat worksheet. This limits the operations that can be done, and the menus change accordingly, as shown in Fig. 13.11f.
If the **Edit ➔ Fixed Rows** option in **Fig. 13.11f** is turned off, you can now edit the labels of the table, as has been done there. If the table is copied to the clipboard, or saved as a “csv” file, with fixed rows turned off, then the labels are copied across. This compares with **Fig. 13.11d**, where just the data were copied.

In **Fig. 13.11g** we show the new **Format ➔ Cells** dialogue for a table without the backend. You can now apply formats overall, and to a selection of your choosing.

Sometimes, particularly with large datasets, the analysis proceeds in “stages”. One example was shown in Section 13.5, where we used daily rainfall data and produced tables of the monthly rainfall totals. We may then wish to use the monthly summary in a further study, perhaps comparing the rainfall with crop data.

For illustration we again use the data from the rice survey, and assume we have the table shown in **Fig. 13.11g**, giving both the number of farmers and their mean yields, for the 4 different villages and the 3 different varieties.

**Fig. 13.11h** Survey data with two variables

<table>
<thead>
<tr>
<th>Village</th>
<th>Variety</th>
<th>Count</th>
<th>Mean Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>SABEEY</td>
<td>NEW</td>
<td>2</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td>OLD</td>
<td>5</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>3</td>
<td>29.5</td>
</tr>
<tr>
<td>KESEN</td>
<td>NEW</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>OLD</td>
<td>3</td>
<td>43.3</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>4</td>
<td>24.9</td>
</tr>
<tr>
<td>NIKO</td>
<td>NEW</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>OLD</td>
<td>2</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>3</td>
<td>26.9</td>
</tr>
<tr>
<td>NANDA</td>
<td>NEW</td>
<td>2</td>
<td>59.7</td>
</tr>
<tr>
<td></td>
<td>OLD</td>
<td>7</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>TRAD</td>
<td>5</td>
<td>36.3</td>
</tr>
<tr>
<td>All</td>
<td>All</td>
<td>36.0</td>
<td>40.6</td>
</tr>
</tbody>
</table>

To save the data from this table back into an Instat worksheet, use the **File ➔ Instat Worksheet** dialogue, shown in **Fig. 13.11i**.
Your main choice is whether to write the summary data to a new worksheet, or to empty columns in the same worksheet as the raw data. We have here chosen the same worksheet. The resulting data looks similar to the 4 columns in Fig. 13.11j. We see that the resulting columns, in the figure below, are of length 12, i.e. there is one row of data for each combination of factor levels. However the margins in the table above have been omitted, because they would mess up any further processing of the data.

These columns could themselves be presented in tabular form. We illustrate by using the Statistics ⇒ Tables ⇒ Summary Tables dialogue, Fig. 13.11k, to produce a table of mean yields, from the data shown in X11 in Fig. 13.11j. The resulting table is shown in Fig. 13.11l. Simple though it is, this table merits further study, because an important principle can be illustrated. Take the mean yield in Fig. 13.11l, for the variety OLD, averaged over the 4 villages. This is 43.7, not the same as the 45.4 that was in the table in Fig. 13.11f.

The key question is “What is our unit?” or equivalently, “What level are we working at?” The data in x12-x15 are at the “village level”, whereas the original data were at the farmer level. There were 17 farmers that used this variety, 5 + 3 + 2 + 7, in the four villages and their mean yield was 45.4.

On the other hand, at the village level, the variety was grown in each of the 4 villages, with means of 49.2, 43.3, 36.1 and 46.4. The mean of these 4 values is 43.7. So, if a village is our “unit” or “level” of interest, then the appropriate mean is 43.7.
If we want to obtain the means, at the farmer level, from the village data, we can use the counts of farmers, in X10, as weights. This is shown in Fig. 13.11m, with the resulting table in Fig. 13.11n.

**Fig. 13.11m** (Table) Data ➔ Weights

**Fig. 13.11n** Results

### 13.12 In conclusion

Overall we hope that this chapter has helped you to see tables in a new light. They are a vital tool in data processing for all types of data. And they do not have to be a dull part of the analysis, nor of a training course.

We believe that the materials in this Chapter, together with the basic notions of data summary (Chapter 12), and of graphics (Chapter 11) can help statistics courses to spend more time on descriptive methods.

This is not to avoid the many concepts concerned with statistical modelling, that have to be introduced, and which we describe in the following Chapters. It is because the ideas in Chapters 11 to 13 are useful in data processing. Often they are most of what is needed, at least for an initial analysis. Also it is easier to explain the uses of statistical modelling if we start from a solid foundation of descriptive methods and ask what more is added by inferential methods.

So what are the limitations of the tables in Instat? One is the size limits of Instat, together with the fact that the tables module can be slow for large datasets. Sometimes survey data can be very large. Instat is limited to 127 columns (i.e. questions), and a few thousand rows, (i.e. people). So for routine analysis of survey data you may wish still to use the industry standards.

One is Excel, where the corresponding facility is called a “pivot-table”. This is a major strength in Excel, and its facilities for tabulation are better than in most statistics packages. The only features in Instat that are not in Excel, are the range of summary statistics offered by the General Tables dialogue, and the facilities for multiple response. Excel can cope with a larger data matrix and also has the easy facility to turn a table to a bar chart.

Among statistics packages the industry standard in SPSS. This has all the facilities shown in this chapter, including those for multiple response data. In addition, a weakness in Instat is its limited text-handling facilities. So the labels for factors often can not be presented as one would like. This can always be remedied after exporting the table to Word, or to another reporting package. SPSS also offers comprehensive facilities for editing the table, so the presentation is just as one would wish.

When a survey is large it may be sensible to organize the structure of the tables initially, perhaps using the pilot study. Then the corresponding commands are saved and the real tables are produced using these commands. We explained the idea of using commands,
rather than just dialogues in Chapter 4 of this guide. This can be done with SPSS. But in Instat, the tables do not use the old server, and this feature is not yet operational.

The full analysis of a survey starts with the type of tables shown in this chapter. Then subsequent analyses might examine relationships as we do in Chapter 17, or include Analysis of Variance, see Chapter 16. One complication is that many surveys are multi-level. The rice survey is an example, where the sampling was village, then field within village. This structure of the data means that the simple standard errors (see Chapter 15) and results from significance tests are not valid. They need to be adjusted for the sampling scheme. This used to require special software, but is now available in some statistics packages, including Stata and SAS.
Chapter 14 - Probability models

14.1 Introduction

Data analysis often involves an underlying probability model for the data. In this Chapter we therefore start the move from purely descriptive statistics and add modelling ideas. In Section 14.2 we describe briefly why this change is often useful and how we can teach the new ideas.

At some stage in analysing a set of data it is usually desirable to compare the distributional properties of the data with properties of the corresponding theoretical distribution. In the remaining sections of this chapter we describe Instat’s facilities to introduce and teach some of the many concepts that are now required. They follow from a single dialogue, shown in Fig. 14.1b, following the menu in Fig. 14.1a. Here we first use the option to draw plots of the different distributions. This option is purely to support teaching. Then we show how the same dialogue is used to provide the significance levels for most of the common tests used in data analysis.

14.2 The role of modelling and inference

We sometimes find reports that “data will be analysed by such-and-such a method”. This sometimes follows from training courses that leap quickly to the teaching of topics such as chi-square tests, or Analysis of Variance.

If we start, instead, from the objectives of the study, then often we find that much of the analysis can be done by producing sensible graphs and tables, using the methods described in the previous chapters. Then a draft report can be written. And it is useful to ensure that trainees realise how much they can still accomplish even if they find difficulties with the ideas to be introduced in this and the remaining chapters.

Given the many concepts that we have to teach, once we move from descriptive statistics to statistical inference, it is useful to pause to explain why descriptive statistics may not be enough. It is also useful to describe how the new ideas may change the contents of the draft report that needed only the simple descriptive methods.

So, the ideas of generalising from a sample to a population, imply sampling ideas and also the ideas of a probability distribution for the population, that we introduce in this chapter.
Then there is the idea, introduced in Chapter 15, of a measure of precision, for estimators, called the “standard error”. This is often included in reports to show the precision with which results apply to the whole population, rather than just being a summary.

So far we have suggested that the tables and graphs in a report are those that correspond to the objectives of the study. But we also want the simplest description that we can give. A simple table or graph may be misleading if we find, usually from fitting a model, that there are important interactions. And a complicated table may be difficult to interpret and unnecessary, if no interactions are present. Thus we often use modelling ideas to refine how the results are presented, i.e. exactly which tables and graphs we will use.

The final preparation is that we urge trainers of non-statisticians to recognise just how many concepts we need to introduce. Descriptive statistics is conceptually easy, while statistical modelling is not. It is for that reason that Instat supplies a variety of teaching graphs, etc. to help ease the considerable pain that this part of the teaching often inflicts.

### 14.3 Looking at the normal distribution

In the dialogue above, accessed through Statistics ⇒ Probability distributions, the simplest plot is of a single normal distribution, shown in Fig. 14.3a, for a mean of 40 and standard deviation of 5. Perhaps more useful, for an introductory discussion, is to plot two distributions. The second graph, in Fig. 14.3b, shows the same distribution, together with the normal distribution of mean 60 and standard deviation 10.

It is just a small step from here to add a plot that includes an important principle in statistics. The mixture, shown from the dialogue in Fig. 14.2c, has a bimodal distribution and also a larger spread than either component. We have also used Edit ⇒ Output ⇒ Presentation Mode so that the lines in Fig. 14.2d are displayed clearly when used for projection, in teaching a large group.
The graph in Fig. 14.3d illustrates three points. First is the importance of considering “structure” when analysing data. If the structure is considered here, then the problem is seen to be a mixture of 2 distributions. If not then “the distribution” is bimodal and the corresponding data are more awkward to process.

A second and related point is that in this example, the mixture has more variability than either component. The idea of a good analysis is to explain as much variability as possible. Here the first step is again to recognise the two distributions if possible.

The third point is the ease with which we can progress from looking at a single distribution, Fig. 14.3a to some interesting topics. For some students this aspect will help to make this part of the subject more relevant and hence easier.

A different option is to include a sample size into the dialogue. This is shown in Fig. 14.3e, for a sample of size 12 and the resulting graph, in Fig. 14.3f, shows the sampling distribution of the mean. The title, in Fig. 14.3f, gives the standard deviation of the sampling distribution, to motivate a discussion of the formula $\sigma/\sqrt{n}$.

Yet another concept is to consider the difference between two normal distributions, often needed in simple inference. As an example, consider the length of the rainy season in a tropical country. Suppose we have found the date of the start of the rains to be approximately normally distributed, with mean 1st May (day 122 in the year) and with a standard deviation of 14 days. Similarly suppose the end is found to have a mean of 1st...
November (day 306) and standard deviation of 7 days. What can be said about the length of the season, which we can write as

\[
\text{Length} = \text{End} - \text{Start}
\]

The dialogue is in Fig. 14.3g and the result is in Fig. 14.3h. There the standard deviation is stated to be 15.6 days, using the result that the variance of the difference is the sum of the variances. There are also options to investigate how this variance is changed, if the two individual distributions are correlated.

**14.4 Looking at other distributions**

The other distributions that can be graphed with this dialogue are the exponential, gamma, extreme value, Weibull, t, \( \chi^2 \), F, binomial and Poisson. The first example, uses the dialogue in Fig. 14.4a, and is the Poisson distribution, with mean 5, together with the Normal approximation. The result is in Fig. 14.4b.

The second, with the dialogue in Fig. 14.4c, shows the exponential distribution, with mean 5, together with the sampling distribution of the mean. In Fig. 14.4d the exact sampling distribution is shown, together with the normal approximation to this distribution. This is therefore an illustration of the central limit theorem.
14.5 Calculating probabilities and percentage points

Probabilities and percentage points can be given for the same set of distributions listed in the previous section. The dialogue in Fig. 14.5a gives the cumulative distribution for the exponential distribution with mean 10. The probabilities are shown in Fig. 14.5b.

With the dialogue as in Fig. 14.5a, the results are also saved to the worksheet, as shown in Fig. 14.5c. This enables a set of tables to be built, for any of the distributions that are in the dialogue. Finally Fig. 14.5d shows a graph of the results.
When calculating percentage points, one small feature we find saves time in our explanations is shown in Fig. 14.5e. When using Student’s t distribution, or the Normal distribution, we often want to know the value(s) that leave 95% of the distribution inside, or equivalently leave 5% outside. The dialogue is shown below, together with the results. Earlier we had to specify the 97.5% point and then start explaining how the value we want leaves 2.5% outside on each side.

This explanation took time and often distracted the non-statistician from our main point that the 95% limits were roughly $\pm 2$, and this was therefore the multiplier from the standard error in calculating confidence limits. Now, although the user sees the 2.5% and 97.5% points in the output, in Fig 14.5f, what they specify is the 95% that they require.
14.6 In conclusion

In both this and the next chapter, the main emphasis in producing these dialogues is to support the teaching of statistics. We invite trainers to review some of the ideas shown here in relation to the way they currently teach those concepts. Our surprise was at how many quite difficult concepts there were, that we had to introduce. This is not a problem for statistics students, but it is a lot to digest for students taking a service course in the subject. No wonder many students find their statistics course to be difficult!

If trainers decide to use the computer as a teaching aid, then they have the choice between a general-purpose tool, such as Instat or special programs that are designed to introduce particular concepts. Such special programs are perhaps more effective for any given concept. Our reason for sometimes using Instat is that we do not have much time for each concept and the use of a specialised program needs to be balanced against the effort of introducing another piece of software.
Chapter 15 - Fitting simple models

15.1 Introduction

We now assume the reader is familiar with the methods for summarising and graphing data, i.e. descriptive statistics. This chapter is used to introduce the concepts of statistical inference. As with the last chapter, on probability distributions, there are key concepts to explain. Here these are the ideas of a standard error, a confidence interval and a significance test.

In this chapter we look mainly at examples from the Statistics ⇒ Simple Models menu, see Fig. 15.1a. We first consider a single sample from a normal distribution in Section 15.2 and then compare 2 normals in Section 15.3. We look at proportions in Section 15.4 partly to emphasise that the concepts apply whatever the type of data. The generality of the ideas is further reinforced in Section 15.5, which considers a simple regression model. This uses the simplest example of the Statistics ⇒ Regression dialogues, as shown in Fig. 15.1b.

These are all situations where there are just a few unknown parameters in the model. The chapter is designed to serve two purposes. The first is that some real problems require these methods. The second is that the concepts apply equally in the more general situations when there are more parameters in the model.

The concepts described here are also covered in the "good-practice" guide called "Confidence and Significance: Key Concepts of Inferential Statistics".

15.2 Confidence intervals for data from a normal distribution

Our first example is the monthly rainfall data from Galle, in Sri Lanka, from 1950 to 1983. The dialogue in Fig. 15.2a gives the results for the data from August, that is in column X9, in Fig. 15.2b.
We also choose a graph that shows the confidence limits and is given in Fig. 15.2c. The graph shows the data plotted in order, which is appropriate here as the values are from the successive years, 1950 to 1983. There seems no obvious trend or other pattern in the data.

The mean is 163 mm with a 95% confidence interval of 134 mm to 192 mm. Students often misinterpret the confidence interval as one that covers most of the data, rather than one that is likely to include the true mean. That misinterpretation is partly because the summary or prediction intervals are also useful. In Fig. 15.2d we have used the same dialogue shown in Fig. 15.2a and included these “summary intervals”, to emphasise the difference. We have also given the graph “by sample” instead of “by observation”.
One source of confusion for trainees is that the mean is sometimes not the parameter of direct interest. For example, in climatology, we may be more interested in estimating particular percentage points of the distribution. An example is shown in the dialogue in Fig. 15.2e where we have estimated the 20% and 80% points of the August totals. An interpretation is that the 80% point is the value that is exceeded just one year in 5, hence it is known as the 5-year return period.

The graph of the results, including confidence limits for the percentage points, is in Fig. 15.2f, and the numerical results are in Fig. 15.2g

The numerical results show that the 80% point is estimated as 230mm with 95% confidence limits from 201mm to 269mm. The eagle-eyed observer may have noticed that this interval is not symmetric. Behind the scenes the calculations use the non-central t distribution and
are not trivial. But this is irrelevant to almost all users, who just need to understand the general concepts of interpreting a confidence interval.

This example is also a good indicator of understanding. If you can explain what is meant by the 95% interval for the 80% point of the distribution, then you really have grasped the idea of confidence intervals!

Finally, in this section, it may be important in teaching to demonstrate that the full data are not needed to derive the results of this chapter. All we need are the “sufficient statistics” and here these are the sample size, the mean and the standard deviation. This can be shown using one of the layout options of this dialogue, as shown in Fig. 15.2h.

**Fig. 15.2h  Summary statistics give same results**

This analysis can then be used to provoke a discussion of the role of the actual data, if this type of model is used. The answer is that the raw data are needed to check on the validity of the assumed model. Here the assumed model is that the data behave as though they are a random sample from a normal distribution. Users often spend too long testing the assumption of normality, when the central limit theorem indicates that this assumption is not crucial, if just the mean is being estimated. However if, as in the example above, other percentage points are being estimated, then the assumption of normality is important.

### 15.3 Comparing 2 samples from normal distributions

For illustration we look at the rice yields from the survey that has been considered in earlier chapters. With two samples, there are alternative layouts for the data. They may be in 2 separate columns, as would be the case with the rainfall totals from the last section, if we wished to compare two different months. Or they may be in a single column, as here, where the yields are all in X6. Then there is a another (factor) column, here X5, which to indicates to which variety each observation belongs. Part of the dialogue is shown in **Fig. 15.3a**, and a graph of the results is in **Fig. 15.3b**. The graph shows the confidence interval for each of the two means separately, and also the confidence interval for the difference. The difference between the two means is 15 cwts/acre with 95% limits from 10 to 20.
The other part of the dialogue is in Fig. 15.3c, which shows that the same dialogue is also used if we wish to compare the variances of the two distributions.

In this analysis, as shown in Fig. 15.3c, we have included the option to test the hypothesis that the two means are equal. The result is shown in Fig. 15.3d and is highly significant. The two means are 30.0 and 45.2 and it is this difference of -15.4 in the means, with a standard error of the difference of 2.42, that has resulted in the t-value of -6.36. This result is also shown in the graph, in Fig. 15.3b, where the difference of zero in the means is seen to be well outside the confidence interval.

The graph indicates that the confidence interval is more useful than the significance test for one and two sample problems. This is because the confidence interval permits any number of significance tests to be done. For example here we find that the 95% confidence interval for the difference between means is between 10.5 and 20.4. The interpretation is roughly that the true difference in the mean yields, is likely to be within this range.

Once we know this, we are pretty certain that the difference is not zero, simply because we think that the difference is between 10 and 20. If a new theory stated that the true difference was 12, then this would be consistent with the data, and we would declare the result non-significant.
And there is a further key point. Knowing the size of the difference is usually more closely connected to the objectives of a study than just knowing that a difference exists. Once we know the size of a difference we can perhaps decide what action to take.

This is the reason that we have made the confidence interval the default result for one and two sample models, with the significance test as an option. We are not against significance tests, though we do find they are much overused.

What is then the role of significance testing? Here, where we are fitting simple models, their role is mainly in training. The concepts of significance testing are also explained in Section 5 of the Inference guide, with the previous sections being devoted to an explanation of standard errors and confidence limits.

Most real problems require inference with more than the simple models described in this chapter. Then the parallel between significance testing and confidence intervals breaks down and the analysis often proceeds in stages. The first stage is usually concerned with finding an appropriate model, and the ideas of significance testing are useful to assist in this choice. We will see many examples in the next chapters. Once we know which model we will use, then we look for the estimates of the parameters, together with a measure of precision (the standard error), so we can proceed.

It is rare that an analysis will need just significance tests, unless the results from the tests are not significant. This is because the significance test is usually of a simple null hypothesis, introduced as a convenient starting point, and not because it corresponds to the objectives of the study. Hence statistical significance usually indicates there is something of interest in the data. Then the objectives of the study will dictate how the analysis should proceed.

15.4 Looking at proportions – an example of a non-normal model

In Chapter 14 we described various probability models, including the normal, exponential, binomial and Poisson models. Data are not always from a normal model and this section symbolises the idea that the concepts of statistical inference apply equally whatever the model. In Fig. 15.4a we show Instat’s menu of simple models again and use the dialogue called “Proportion, one sample”

We use the data from the rice survey again, this time looking at the variety of rice grown. This is an example of data where the farmer gave one of 3 responses, namely “New”, “Old” or “Traditional” rice. This type of qualitative or categorical data is very common. We will look at other examples in Chapter 18 of this guide.

We may have “ordered” categories, like disease score on a 1 to 9 scale, or a 5 point scale for acceptance of a product, ranging from “very good” to “very bad”. Sometimes we have “sort-of-ordered” data, like the 3 varieties in this survey. Here we might think of the ordering going from “New” to “Old” to “Trad” once we look at the mean yields, but perhaps if we had looked at taste, or disease resistance the order would have been different.
Descriptive analysis of data of this type was shown in Chapters 12 and 13, where we looked at tabulation. Here, as an example, Fig. 15.4b shows that 15 out of the 36 farmers, i.e. 42% grew traditional rice. This is the information we will examine further in this section.

One issue that often causes confusion in training courses is the extent to which the data that we use have already been summarised. We can start with the raw data, as indicated in Fig. 15.4c. Then we define the code 3, corresponding to “Trad” as a success and simply count the number of times a “Trad” occurred. So this is just like having a column of 0’s and 1’s, or “No” and “Yes” and we count the number of 1’s or “Yes”. The data are then from what is called a Bernoulli model, i.e. each observation is either a success or a failure.

The alternative is to consider the summary data, as shown in Fig. 15.4d as an observation from the binomial model, with n=36 and r = 15. This is the equivalent for the normal model of giving just the summary statistics, as was shown earlier in Fig. 15.2e.

Training courses still often use just the summary data when showing the analysis of this type of problem and we feel that more use of the raw data would help users to see the simplicity of the subject.
The formulae and methods for calculating confidence intervals are given in Armitage and Berry (1994). Instat provides 3 options, of which two are shown in Fig. 13.4e above. The exact intervals calculate sums of binomial probabilities, (the result is related to the calculation of a cumulative probability from a corresponding F distribution), and here we see that the 95% interval ranges from 0.25 to 0.6. The simple normal approximation calculates the standard error as \( \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.417 \times 0.583}{36}} = 0.082 \) here, and then calculates

\[
0.417 \pm 1.96 \times 0.082 = 0.256 \text{ to } 0.578
\]

The width of this interval indicates that 36 observation is a small sample, if only “Yes” or “No” is to be recorded for each individual. We return to this point in Chapter 21 when looking at the calculation of sample sizes.

15.5 Simple regression – emphasising the general concepts

Chapter 17 is devoted to regression models in general. Here we use this third situation mainly to emphasise that the concepts of statistical inference apply generally. The first two models were normal, in Sections 15.2 and 15.3, and binomial in Section 15.4. For illustration we again use the survey data and look at the relation between yields and fertilizer. The plot, with the fitted line is shown in Fig. 15.5a.

The simple regression dialogue has the same defaults and options for confidence intervals and significance tests that we have seen earlier. Here the obvious test is that the slope of the true line is zero, because that would indicate there is no (linear) relationship between yield and fertilizer.
The results are shown in Fig. 15.5c, and show the four components, namely the fitted equation, the standard error, confidence limits and the significance test.

We see again, as described for the simple normal models in Section 15.3, that the results of the significance test are obvious once we know the confidence limits for the slope. If, as shown by the 95% limits, the true slope is likely to be between 6.4 and 11.5, then it is hardly likely to be zero!

We believe there is a lot to be gained from using this type of example to link the concepts of simple inference, described earlier in this chapter, with the ideas of fitting more complex models, that will be discussed in succeeding chapters. The first is that the model for simple linear regression is here

\[
\text{Yield} = \alpha + \beta \times \text{fert} + \text{residual}
\]

We estimate the \( \alpha \) and \( \beta \) with \( a = 27.66 \) and \( b = 8.948 \) to give the fitted equation of

\[
\text{Yield} = 27.66 + 8.948 \times \text{fert}
\]

In general we like to think of

\[
\text{Data} = \text{Pattern} + \text{Residual}
\]

This idea applies equally to the other models that we have used in this chapter, but they are not so obviously formulated in this way. Also it is very easy to see that this simple straight-
line model may not be all that we need. Perhaps the “pattern” we need a quadratic.
Perhaps we also might consider a model that includes the varieties in some way, as well as
the fertilizer.

Later in a training course, one additional bonus is that this simple regression model can also
be used to introduce the idea of an Analysis of Variance table. This is an option on this
dialogue, as shown in Fig. 15.5b. If used, for the same example, it gives the results shown
in Fig. 15.5d.

**Fig. 15.5d Analysis of variance (ANOVA) table**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2993.7</td>
<td></td>
<td>51.90</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>34</td>
<td>1361.04</td>
<td>57.678</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>4954.74</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-squared = 0.6042

To understand the role of an ANOVA table, it is useful to look for the parallels and additional
information that the ANOVA provides, for this simple model. The F value of 51.90 is simply
the square of the t-value, which we saw earlier was 7.2. The 34 degrees of freedom is
because our model has 2 parameters, the intercept and slope. And we saw the 34 in the
description in Fig. 15.5c earlier. Finally the 57.678 is the residual mean square, giving a
residual s of √57.678 = 7.595. This indicates the variation that our model cannot explain,
and we can see this variation in the graph in Fig. 15.5a.

Once the terms have been explained, the key point is that the ANOVA table is no more use
here than the significance test of the slope, shown in Fig. 15.5c. And that is not as useful as
the confidence limits. This is because this is “simple” linear regression, where the
regression model only has one degree of freedom, and hence the confidence limits for the
slope provide more than the significance test. However, the method of ANOVA sets the
scene for later chapters where more parameters will be estimated.

Hence, if a basic training course is covering no more than simple linear regression, then
there is little point in introducing the concept of an ANOVA table. For courses that go
further, if trainees could understand the components of an ANOVA table, for this simple
model, then the role of an ANOVA table should be clearer, for more complicated models,
when it is needed. To support this understanding, Instat offers a series of “teaching plots”,
as shown in Figs. 15.5 e, f, g, h. If these are the only active plots, then they tile, roughly as
shown here.

**Fig. 15.5e** shows the 2 alternative models. The line parallel to the x-axis is the “null” model,
and the other is the fitted regression line. In **Fig. 15.5f** we see that the (regression) model
explains 2994 of this sum of squares, which is good value for just one degree of freedom,
given the total, in **Fig. 15.5d**, of 4955.
The idea is that a good model will explain as much as possible of the variation of the data, leaving the residual sum of squares to be as small as possible. **Fig. 15.5g** and **Fig. 15.5h** have the same scale so the difference in sum of squares can be seen on the graph, as well as read from the title.

In Chapter 16, Section 16.2, we will see 4 similar graphs to show the effect of using the variety factor, rather than the fertiliser, in the model. That model has 2 degrees of freedom and its residual sum of squares is lower, at 1427. Then, in Chapter 17, Section 17.4, we will see that both the fertilizer and the variety can be combined into a model with 3 degrees of freedom, leaving a residual sum of squares of just 732.
15.6 In conclusion

There is much that we could have added to this chapter that is also important. For example, the 2-sample dialogues, for both the normal and binomial models allow a “paired” option that can be used to introduce the value of blocking. Often the pairing is a person who may be tasting different foods, say, or trying different cosmetic treatments on each cheek. Then this idea can also be used to introduce the idea of data at multiple “levels”, i.e. between and within person.

The dialogues for the normal models also include confidence intervals and tests for variances, as well as the means. As well as introducing the F-distribution, this also provides an opportunity to look at the assumptions of a standard analysis.

The examination of the adequacy of the assumed model is another key area. To take just one point, we find that trainees often think that an assumption for many analyses is that “the data are from a normal distribution”. This is a sensible view for the simplest models, particularly when looking at a sample from a single normal distribution. Hence this dialogue provides an option to produce a normal probability plot.

However, once we have understood the general idea of

\[
\text{Data} = \text{Pattern} + \text{Residual}
\]

then it is the residuals, that we must use to test the assumptions of the model, rather than the “data”. This is easy to illustrate with the simple regression model, introduced in Section 15.5.

Our claim is that once the concepts introduced here, and in Chapter 14, have been understood, then students have the solid foundation in statistical methods that can support all their further work.

The large number of concepts that have to be introduced was a surprise to us. No wonder many students find the subject to be confusing!

When teaching postgraduate students or in-service courses, the topics introduced in these last two chapters are usually what we would like to assume. Often we cannot, and courses are made less effective, because some participants did not really understand the ideas on standard errors or confidence limits. In such courses, what we often do now is first to give each participant the short guide on

“Confidence and Significance: Key Concepts of Inferential Statistics”.

Then we either allow for a short review session within the first day, or we timetable up to half a day to re-cover those ideas.

If, instead of using Instat, you wish to use Excel, the SSC-Stat add-in includes similar dialogues for normal and binomial models. They are intended primarily to support the concepts of simple inference.

Other statistics packages provide similar dialogues.
Part 4 – Further Methods

In the final chapters we are mainly concerned with data analysis where models have more parameters than those we considered in Chapter 15.

In Chapter 16 we describe models where the Analysis of Variance is used as a tool in the analysis, while regression models are considered in Chapter 17. We look at the analysis of qualitative data in Chapter 18 and models where time is a component in Chapter 19.

In Chapter 20 we describe how to use the computer to simulate data. In Chapter 21 we return to the beginning and look at some of the roles of the computer in supporting the design of data collection activities.
Chapter 16 - Analysis of variance

16.1 Introduction

In Chapter 15 we introduced the idea of a model, expressed simply as

\[ \text{Data} = \text{Pattern} + \text{Residual} \]

This chapter is concerned with models where the pattern is mainly caused by “factor” or “category” columns, such as the blocks or treatments. The next chapter is on regression modelling and deals with models where the pattern is mainly caused by variates.

Data from designed experiments are usually analysed by analysis of variance techniques. These methods are also useful for the analysis of other types of data, for instance from surveys. One difference is that surveys usually give rise to unbalanced “designs”, while many experiments permit a simpler analysis.

On-farm trials are often between experiments and surveys, and show some characteristics of each.

The simple example described in Chapter 3, and shown in Fig. 16.1a indicates the three components of any experiment. The first component is the “layout” of the trial, here an experiment in simple blocks, given by X2. We apply “treatments” that correspond to the objectives of our study. Here there are 3 treatments, shown in X3. Then we take measurements. Here there is just one measurement, given by the data in X1.

Our analysis is then of the measurements, with the structure of the data given by the layout and treatment factors. In this Chapter, we show the use of Instat’s three main dialogues, shown in Fig. 16.1b, labelled “One-way”, “Orthogonal” and “General”. The General dialogue uses an algorithm that was previously in separate software called Genanova. A special guide called “The Statistical Background to ANOVA is available, that describes many of the concepts concerned with the analysis of experimental data.


16.2 One-way ANOVA

As in previous chapters we use the data from the rice survey, and here examine the extent to which the yields are different for the 3 varieties. The dialogue is shown in Fig. 16.2a, with the output in Fig. 16.2b. The output provides the Analysis of Variance table, the treatment means and their standard errors.

There are various graphics options. The two main options are to examine the fitted model, i.e. the “pattern”, which is here the means, and the second is to look at the residuals. These options are available for all the ANOVA and regression dialogues. We first look at a special option, called ANOVA in the dialogue above, which is designed to support the teaching of the Analysis of Variance. This produces the 4 graphs shown in Figs. 16.2c, d, e and f.

These graphs are designed to assist in the explanation of the terms, degrees-of-freedom and the sums of squares in an ANOVA table. This set of graphs is the parallel of those described in Section 15.5, when fitting a simple regression model to the same set of data.

Here the model, with 2 degrees of freedom, explains 3528 of the total sum of squares, leaving a residual sum of squares of 1427. These are the same sum of squares as are given in the ANOVA table, Fig. 16.2b.
In Chapter 15 we argued that significance testing is hardly needed once the ideas of confidence intervals are understood. This is no longer the case. With 3 treatments, this is the simplest example where the F-test, that is part of the ANOVA table, fulfils a separate role. It tests the overall hypothesis that all the true means are equal, against the alternative that there are differences between them. The result in Fig. 16.2b is highly significant and this provides a “passport” to examine these differences.

It is often informative to graph the residuals, to see if the chosen model is adequate. Here we already saw, in Section 15.5, that the yields are related to the fertiliser and this is confirmed if we plot the residuals from this model, which has fitted the varieties, against the fertiliser levels. It looks as though we should include both fertiliser and variety in the model together. We cannot proceed further in this chapter, but wait for Chapter 17, Section 17.4 to see the resulting model.

When analysing experimental data, our main aim is usually to compare the treatments. Here we see from Fig. 16.2b that the means are given as 59.6, 45.4 and 30.0, together with the three standard errors. We see again that the idea of a standard error, introduced in Chapter 15, remains important for these more general models.

However if our main aim is to compare the means, then we really want the standard error of the difference between the means. We see how this is done in general, in Section 16.8, when we consider the subject of treatment contrasts. Here we continue the connection with
Chapter 15 by using again the Simple Models ⇒ Normal, Two Samples dialogue, which has a special “following ANOVA option, shown in Fig. 16.2i.

Fig. 16.2i Stats ⇒ Simple Models ⇒ Normal, Two Samples

If the dialogue is used straight after an ANOVA, then some of the terms in Fig. 16.2i are completed automatically. If not, then the residual mean square and degrees of freedom should be taken from the ANOVA shown in Fig. 16.2b. As an example we compare the means from the new and old varieties. The means and counts are taken from Fig. 16.2b and typed into the dialogue. Then we proceed as shown in Chapter 15, giving the results shown in Fig. 16.2j. The results can also be displayed graphically as was shown in Chapter 15.

16.3 Including blocks in the model

In many textbooks it is conventional to progress from the one-way ANOVA to the complete randomised block design. This approach keeps the arithmetic simple, but can distract the student from understanding the real use of blocking. Blocks are mainly to reduce the unexplained, or residual variability in the data. Here we introduce blocking in a general context.

As an example, we continue, from Section 16.2, where we examined the yields in the survey, in relation to the variety of rice. Now we allow for the different villages as a blocking factor. Our model is therefore that 

\[ \text{Yield} = \text{Village effect} + \text{Variety effect} + \text{Residual} \]

Because this was a survey, the varieties are unlikely to have been used equally in the different villages. We can see this with the simple table of counts, shown in Fig. 16.3a. Had this been a trial of the same size, we might have arranged for 3 of the 36 observations to have each variety/village combination and Fig. 16.3a shows the actual replication is nothing like this. For example, we have 4 farmers who grew the new improved varieties and they were from only 2 of the villages.

Our particular interest is in the treatment means. In the last section we saw in Fig. 16.2b, and below in Fig. 16.3b that these means are 59.6, 45.4 and 30.0 for the 3 varieties. But that was from a model that did not allow for a possible village effect. Here we are fitting a model that does include the villages in the structure.
16.4 Simple use of the orthogonal ANOVA dialogue

We now return to the simple randomised block data, shown in Fig. 16.1a earlier. Part of the Analysis of Variance ⇒ Orthogonal dialogue is shown in Fig. 16.4a. In the dialogue, we have chosen to give a simple graph of the data, with the treatment means and this is shown in Fig. 16.4b.

In this dialogue, the order is relevant when specifying the factors. This order of the factors is the same as it will be in the resulting ANOVA table, shown in Fig. 16.4c. Here it does not matter, but it is conventional to specify the blocks before the treatments. This order will matter more in subsequent sections, when the treatment structure is more complicated.

By comparison with Fig. 16.2b we see here that the residual sum of squares is lower at 995, compared with 1427 using the model without the villages. One other change is that the variety means have been adjusted for the village effects, and this should give a “fairer” comparison of the varieties.

We continue this example in Section 16.8 where we look in more detail at the comparison of the treatment means.
The output in Fig. 16.4c gives the ANOVA table, followed by some simple diagnostics that are described in Section 16.5. Then there are the tables of means together with standard errors.

**Fig. 16.4c Results**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>3</td>
<td>2330.3</td>
<td>776.75</td>
<td>2.0</td>
<td>0.214</td>
</tr>
<tr>
<td>Treat</td>
<td>2</td>
<td>4212.5</td>
<td>2106.3</td>
<td>5.4</td>
<td>0.045</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2321.5</td>
<td>386.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>6064.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall Mean = 332.8  s (Residual) = 19.67

Coefficient of Variation = 5.9 

The following observations have large residuals:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Value</th>
<th>Residual</th>
<th>se</th>
<th>Ratio</th>
<th>%RSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>313</td>
<td>20.59</td>
<td>13.9</td>
<td>1.48</td>
<td>18.2</td>
</tr>
<tr>
<td>11</td>
<td>373</td>
<td>25</td>
<td>13.9</td>
<td>1.87</td>
<td>29.1</td>
</tr>
<tr>
<td>12</td>
<td>302</td>
<td>-26.67</td>
<td>13.9</td>
<td>-1.92</td>
<td>30.6</td>
</tr>
</tbody>
</table>

**MAIN EFFECTS**

<table>
<thead>
<tr>
<th>Block</th>
<th>Treat</th>
<th>Level</th>
<th>Mean</th>
<th>Level</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treat</td>
<td>0</td>
<td>353.7</td>
<td>306.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>E</td>
<td>321.7</td>
<td>349</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>F</td>
<td>337</td>
<td>342.6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>4</td>
<td>318.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SE acan = 11.36  SE acan = 9.835

SE diff = 16.06  SE diff = 13.91

We return to this example in Section 16.8, when we look in more detail at the treatment means.
16.5 ANOVA diagnostics

The most common diagnostic of the adequacy of an Analysis of Variance is the Coefficient of Variation (cv). This is the residual standard deviation as a percentage of the overall mean, e.g. $100 \times 19.67/332.8 = 5.9\%$ in Fig. 16.4c. If the cv is "too large" it may indicate some problem with the data or with the model being fitted. What is considered "too large" depends on the type of variable being measured. For yield variables in agronomic trials, 20% is often considered acceptable, while 50% would be too large. The cv is sometimes useful, but it is a very overused statistic. It is often more useful to examine its two components. If the cv is surprisingly high, it may be because the overall mean is low, or because the residual standard deviation is high, or both. The results in Fig. 16.4c therefore also give the two components of the cv.

The cv is not the only diagnostic that should be used. It is always a good idea to look at the individual residuals to check the adequacy of the model that you have fitted. Some software, textbooks and hence training courses overlook this step, although they may make quite an issue of examining residuals in regression. In both ANOVA and regression, however, we are fitting a linear model to data by the method of least squares with the same underlying assumptions, and these assumptions need to be checked. The ANOVA dialogue automatically notes any observations that have a "large" residual. For example in Fig. 16.4c, observation 12 is noted as contributing 30% to the residual sum of squares. This does not necessarily mean that there is a problem with this observation, but it provides an initial pointer to which observations to examine first, if there are thought to be problems with the data or the model.

Residual plots can also be given. The dialogue for the same example used in Section 16.4 is shown in Fig. 16.5a, and provides for 3 different options to examine the residuals. These are the same for all the ANOVA dialogues. Here we show the plot of the residuals against the fitted values in Fig. 16.5b. There is no obvious pattern in these residuals.
16.6 Factorial treatment structure

In Fig. 16.6a we show the data from an experiment with 8 treatments. This was a trial on water uptake in frogs and toads from Mead, Curnow and Hasted, (1993), pages 114-118. The treatments consisted of 3 factors, each at 2 levels, commonly called a 2*2*2 factorial treatment structure.

The treatment structure may be clearer from the display of the means, given in the table shown in Fig. 16.6b. We used the Statistics ⇒ Tables ⇒ Summary dialogue described in Section 13.3 to provide this table.

The last column of the table of means in Fig. 16.6b indicates that there may be an “interaction” between two of the factors, species (S) and hormone (H). We see that for one species, toads, the hormone mean is 24.9, much more than the control mean of 10.9. In the other species, frogs, there is almost no difference between the two means. So the effect of the hormone seems to depend on the species, i.e. there appears to be a hormone-species “interaction”.

In Instat we need to derive columns to model the interaction, before proceeding to the analysis. This step uses the Manage ⇒ Data ⇒ Interactions dialogue as shown in Fig. 16.6c. It only has to be done once, because it adds the 4 interaction columns to the worksheet. So, if we wish to re-analyse the data on a future occasion we can proceed directly to the second step.
This second step is to use the Statistics ⇒ Analysis of Variance ⇒ Orthogonal dialogue, which is completed as shown in Fig. 16.6d. The order of the factors and interactions in this dialogue, here X2-X8, will correspond to the successive rows in the ANOVA table.

Some of the results are shown in Fig. 16.6e. We see, from the ANOVA table, that all 3 main effects seem important and there is an indication of an interaction, between species and hormone level. These means were all given earlier, see Fig. 16.6b.

As a second example we look briefly at an experiment that includes simple confounding. Cochran and Cox (1957) describe a fertiliser experiment on the yield of beans, with 2 replicates and 16 treatments. Each replicate has been divided into 2 blocks, each of 8 plots. So there are two blocks for each replicate. On the treatment side, this is a 2*2*2*2 or 2^4 treatment structure, with the 4 treatment factors being Dung (organic fertiliser) and N, P and K. Part of the data are shown in Fig. 16.6g.
As is indicated in Fig. 16.6h and Fig. 16.6i the 2 steps in the analysis are the same as for the ordinary trial, that does not have confounding. The only difference is that the 4-way interaction term that is in X17, is not included in the treatment structure, in Fig. 16.6i, because its effect is already in the block term.

In fact, even if the experimenter forgot to omit the confounded effect in the analysis, the software will sometimes be forgiving. We show the ANOVA table in Fig. 16.6j, for the situation where X2-X17 (instead of X2-X16) was specified as the treatment structure.

The results indicate that the only effects of importance are an effect of Nitrogen and an interaction between dung and P. We used the tabulation dialogue to show the interaction table in Fig. 16.6k, where we see that the message is to apply either neither, or both.
We have emphasised, in Section 16.3, that the main role of blocking is to improve the precision of experiments. Hence the concept of "confounding" is important, because it allows the two key concepts of factorial treatment structure and small blocks to be combined. We hope that this demonstration of the ease of analysis of such trials will tempt trainers to include this topic in their courses and experimenters to consider this feature in future designs. It is often ignored, because it is a late chapter in a textbook and hence appears to be a complicated idea.

16.7 A split plot experiment

In Fig. 16.7a we show some of the data from a split plot experiment on the yield of lettuce (Mead, Curnow & Hasted [1993], pp 133-136). There were 4 blocks. The mainplot treatments were three uncovering dates. They were split into six subplots for each of six varieties.

In this analysis, the main plot error term is the interaction between blocks and the uncovering dates. We first use the Manage ⇒ Data ⇒ Interactions dialogue to get this column, Fig. 16.7b, and also the interaction between the treatment terms, Fig. 16.7c.
The ANOVA dialogue is shown in Fig. 16.7d. The factors and interactions in the model must be given in the order that they will be presented in the ANOVA table. Then X5, the block by date interaction, is specified as the main plot error term.

Fig. 16.7e Interaction plot

The plot, shown in Fig. 16.7e, indicates that there is an interaction between the two treatment factors. This is confirmed by the ANOVA table, shown in Fig. 16.7f. The means and standard errors are also given in the output, though not shown here.

The two-way table of means, Fig. 16.7g, is given using the tabulation facilities, described in Chapter 13.
To complete this example, we quote the summary of the results from Mead, Curnow, and Hasted (1993) page 135:

"Differences between varieties varied with uncovering date. For uncovering date x, varieties D and E outyielded all other varieties significantly; for y, differences between varieties were small; for z, varieties D, E and F gave significantly higher yields than varieties A, B and C. Although the mean yield declined between y and z for 4 of the 6 varieties the decline was significant for variety C only. (All significance statements refer to the 5% significance level.)"

It is also possible to analyse more complicated designs, such as split, split plots, or strip plot designs. However, a brief warning on design. Split and strip plot designs are important, but they are also overused. As a general guideline, use them only if some factors need large plots, while others do not. Otherwise use the randomised block design. If the blocks are too large consider using the technique called "confounding", an example of which was given in Section 16.6.

### 16.8 Examining the treatment means

The basic use of the Analysis of Variance dialogues gives an ANOVA table, followed by tables of means. In this section we see how to examine the treatment means in more detail. There are many alternatives, and which you choose should depend on the objectives of the study. The alternatives include:

- Look at the tables or graphs of the means, together with the standard errors, as they are sufficient to write the report. An example is shown in Section 16.7, see Fig. 16.7f and Fig. 16.7g, and the conclusions that follow.

- Look at particular treatment comparisons, either from the ANOVA output, or using the Simple models ⇒ Normal Two Sample dialogue with the “following ANOVA option. (See section 16.2, Fig. 16.2i for an example.)

- Define appropriate “contrasts” that correspond to the precise objectives and include them in the analysis.

Here we will mainly describe the facilities to include “contrasts” in the analysis.
With the simple experiment shown in Chapter 3, and considered in Section 16.4, there were 3 treatments. The first is the control while the others are 2 test treatments. Hence the 2 degrees of freedom for treatments could usefully be split into separate components. We first enter the appropriate coefficients into the columns in the worksheet, with one column for each contrast. This is shown in Fig. 16.8a.

These contrast columns are the same length as the number of levels of the treatment factor. So here they are of length 3. The name “contrast” is appropriate, for example X4 is contrasting the control with the other two treatments. So the sum of the coefficients of each column is zero.

Fig. 16.8a Defining contrasts

Contrasts may be used either with the Analysis of Variance ⇒ Orthogonal or General dialogues. Their use with the former is shown in Fig. 16.8b, with the extra output shown in Fig. 16.8c.

Fig. 16.8b Using contrasts

Fig. 16.8c Results including contrasts

<table>
<thead>
<tr>
<th>ANOVA TABLE for Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Block</td>
</tr>
<tr>
<td>Treat</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAIN EFFECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Level</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

| SE mean | SE mean | 9.835 |
| SE diff | 15.06 | 13.91 |

<table>
<thead>
<tr>
<th>CONTRASTS for factor: Treat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
</tr>
<tr>
<td>X4OE</td>
</tr>
<tr>
<td>X4E</td>
</tr>
</tbody>
</table>
In Fig. 16.8c the extra output shows the sums of squares for the 2 contrasts is 4134 and 78. These add to the treatment sum of squares, given above, of 4212, so we see that most of the treatment sum of squares can be attributed to the first contrast, and this is confirmed by the corresponding F values.

We may also compare the values of the contrasts with the means, given in Fig. 16.8c. They are calculated using the coefficients, so the first contrast is

\[0.5 \times (349 + 342.8) - 306 = 39.375\]

with standard error of 12.0

The second contrast is just a difference between 2 means, so its standard error, of 13.91, is the same as is given with the ordinary results in Fig. 16.8c.

These contrasts are called “user defined”, because we have defined them specially to correspond to our reasons for choosing these treatments. An alternative is provided in the second example. In Fig. 16.8d we show some of the data from Mead, Curnow and Hasted (1993) page 106.

This is an investigation on the survival of Salmonella in an experiment where there are 3 levels of sorbic acid and six levels of water activity. There were 3 replicates, of the 18 treatments and a randomised block design was used. The data for analysis were log(density/ml) measured 7 days after the treatments were applied.

To prepare for the analysis we have already used Manage ⇒ Data ⇒ Interactions to define X6 as the interaction between Density and Water activity. The graphs of the means from the ANOVA are useful, and are shown in Fig. 16.8f and Fig. 16.8g, below. The first shows the result, from the full model, because there was an indication of an interaction. The graph of the model without the interaction is given also, partly because the interaction effect is small, but also as a teaching point. We see that no interaction implies parallel curves.
Here the 6 levels of water activity change equally from 98 to 78. It might be of interest to see if the large effect of water activity could be modelled by a straight line, (unlikely, given the graphs above!) or perhaps a simple polynomial.

We first put the levels of this factor in a column and then use the Manage ⇒ Data ⇒ Polynomials dialogue, as shown in Fig. 16.8h. We choose orthogonal polynomials, because they will make it easier to interpret the results. Here the factor levels were equally spaced, but this method works equally if not.

This step has generated three columns as shown in Fig. 16.8h, and each adds to zero, so they can be used as contrasts. The dialogue for the analysis is shown in Fig. 16.8j and some of the output is in Fig. 16.8k. We see, from the ANOVA table, that most of the variability in this study is due to the different levels of water activity. And the contrasts indicate that a quadratic model will be adequate.
Some readers will be disappointed that they are reaching the end of this section and there
has been no mention of “standard” comparison of means routines, such as multiple range
tests. In Section 16.10 we explain why.
16.9 Coping with common problems

The software must be able to cope easily with common complications, simply because they are common! We first discuss missing values. Then we consider what can be done, such as transforming data, if the assumptions of a standard ANOVA are not satisfied.

Missing factor or data values are simply ignored in the One-way and in the General dialogues. It does not matter how many you have, because the analysis does not assume that there is equal replication.

If you are using the orthogonal ANOVA dialogue then you can not have missing values in the factor levels, but some data values may be missing, and therefore entered as * in the worksheet. The dialogue is unchanged, because Instat detects the presence of the missing values and analyses the data accordingly. The results are then approximate and therefore should be treated with caution. If more than one or two observations are missing, the data are usually sufficiently non-orthogonal that the analysis should be handled with the General dialogue, described in Section 16.3, or with the general linear models dialogue, described in Chapter 17.

With the orthogonal ANOVA dialogue, the method of adjusting for missing values is iterative, and is as follows. The missing values are first estimated as the overall mean. The data are then analysed. The residual from each missing value is then subtracted from the original estimate and the analysis is repeated. This is continued until the residuals for each missing value are close to zero. Four important points are as follows:-

i) The residual mean square is correct. If a non-orthogonal analysis is performed the same value should result.

ii) The total sum of squares was correct on the first iteration (i.e. originally the missing values were put equal to the overall mean and hence contributed nothing to the total sum of squares). The amount by which the total sum of squares has increased is therefore some measure of non-orthogonality. It is given with the results as an "Index of orthogonality" - 100% is orthogonal and it is larger when there are missing values.

iii) One degree of freedom is subtracted from the residual for each missing observation. The algorithm does not detect situations where all observations for a particular treatment combination are missing.

iv) The standard errors for comparing treatment means are not adjusted for the different sample sizes caused by the missing observations, or for the non-orthogonality.

When the assumptions for an ANOVA are not satisfied, then one possibility is to transform the data. This can either be with the Manage ⇒ Calculate or the Manage ⇒ Transformations ⇒ One to One dialogues, both of which were described in Chapter 8. The transformation may alternatively be typed as a command in the Commands/Output window. For example, when the data are counts, then a square root transformation, for example

: x9 = sqr(x1+0.5)

could be given. Then x9 is analysed, rather than x1.

If data have been transformed for analysis then one minor problem is how to present the results. The problem is that the information is often easier for the reader, on the ordinary scale, but standard errors of comparisons apply on the transformed scale. It is usual to present both, as described in Mead, Curnow and Hasted (1993), page 158. Within Instat the dialogues all have a "Save" option, to save the fitted means back to the worksheet. Then, if needed, the Manage ⇒ Transformations ⇒ One to One dialogue has an "Inverse" option, to give the reverse transformation for some common situations.
When data are counts or percentages, there is a theoretical background to the required transformation. A modern alternative is to analyse the data as an example of a generalised linear model. This provides the same flexibility of analysis for counts from a Poisson model or data from a binomial model, plus others. It requires a more powerful package than Instat. But with the ease of use of modern statistical software, this type of analysis is possible by non-statisticians and we strongly recommend that researchers investigate the possibility.

In other situations, transformations are often used because some treatments appear to be more variable than others. One of the key assumptions of the ANOVA analysis, described in this chapter, is that in our model,

\[ \text{data} = \text{pattern} + \text{residual}, \]

that the residuals have the same variance, whatever the treatment or block. If this is not the case, then transformations can be used. Sometimes it may be appropriate to transform the data and do the analysis, just to see whether the conclusions are changed. It often adds a “comfort factor” if the conclusions are the same, whatever the scale.

Sometimes a simpler alternative to a transformation is to consider the data in two or more parts. As an example, we consider the split plot experiment analysed in Section 16.7. Here there were 4 blocks, 3 uncovering dates and 6 varieties of lettuce. In such an example, if two varieties had very different yields to the others, then it may be appropriate to consider these varieties in a separate analysis.

For illustration, in this example we note that the 4th block has a lower mean than the other three and try an analysis omitting this block. This is done using the Manage ⇒ Transformations ⇒ Select dialogue as shown in Fig. 16.9a, followed by the appropriate analysis of variance for the subset as shown in Fig. 16.9b.

In conclusion

Instat is designed to support the teaching of statistics so we begin this final section by emphasising how easy it is now to teach aspects of a non-orthogonal analysis. In Section 16.3 we considered the analysis of the survey data to assess the differences between the varieties of rice, allowing for village effects as a blocking factor. There are 3 varieties and hence 2 degrees of freedom for this component. We now continue that analysis and look at 2 contrasts, shown in Fig 16.10a. One is to compare the control with the improved varieties and the other to compare the two improved varieties with each other.
The second contrast is a comparison between 2 (adjusted) means, i.e. $57.15 - 45.14 = 12.01$. We see from the results in Fig. 16.10c that the “efficiency” of this contrast is 0.959 and the teaching point concerns what exactly is meant by this efficiency.

We saw earlier that the villages and varieties are non-orthogonal. So we have gained in precision by allowing for the villages, but lost in efficiency. When estimating this contrast the efficiency is 0.959, while it would be 1 if our design were orthogonal. To help explain what “efficiency” is, the Simple Models \(\Rightarrow\) Normal Two Samples dialogue has a special option to compare means, following this general ANOVA. It is completed as shown above. Part of the results are shown below. We see, apart from rounding error, that the standard error of the difference is the same from the simple models dialogue as from the general ANOVA. It is now calculated simply as

$$s \sqrt{(1/n_1 + 1/n_2)/\text{eff}} = 5.76 \sqrt{(1/4 + 1/17)/0.959} = 3.2688$$
So, if trainees have understood the simple inference ideas covered in Chapter 15, these concepts can be extended to provide insights into non-orthogonal analyses.

We consider the facilities in this chapter to be one of the strengths of Instat. However we now explain why multiple comparison of means have not been included. Briefly there are 2 reasons:

- They almost never correspond to the objectives that led to the experiment. If they do not relate to the objectives, then there is no point in adding them to the analysis.
- They often distract users from doing a more appropriate analysis. This may involve contrasts, see Section 16.8, or a graphical display of the means.

We would argue further that the insistence on such tests is related to the overuse of significance tests in statistics generally, as we outline in Chapter 15. We describe our views on multiple comparison procedures in more detail in the good-practice guides on Inference and on the Presentation of results.

Instat is intended to complement the existing statistics packages. Most do include a plethora of multiple comparison procedures. We consider this to be one reason why experimental data are often so poorly analysed. Hence, if their omission here causes users of those packages to analyse their data in a more thoughtful manner, then Instat will have done its job!

With one honourable exception, this area of ANOVA is one that is relatively poorly catered for in many statistics packages. Some would argue that there is little need to have a special section for this type of analysis, because it can all be considered within the more general regression ideas that are the subject of Chapter 17.

We believe it remains useful to have a simple facility for the analysis of standard experiments of the type that remain popular in agricultural and many other research areas. Genstat, is the only package that provides really comprehensive facilities that makes the analysis of all types of experimental data into an easy procedure. This is partly because of its comprehensive use of “factors” as a particular type of column.

So what might be useful that is not in Instat, but is handled by the ANOVA dialogue in Genstat. Here is a partial list:

- Factorial structure, plus a control, analysed neatly into its components.
- Multi-way tables of means shown in the standard ANOVA output in tabular form.
• Split-plot analyses with all standard errors of differences presented. (This is in Instat for simple split plot, but not split-split, etc).

• Analyses of balanced and partially balanced designs, like lattices, within this framework. Also combination of information over different levels for these designs.

• Contrasts given automatically for interactions, if defined for main effects.

• Analysis of covariance possible for any number of covariates. (Up to 2 can be included in Instat’s general ANOVA dialogue.)

We consider the need for more advanced facilities at the end of Chapter 17. For example a split-plot design has information at multiple levels and the general processing of multi-level data is beyond the capabilities of Instat. Similarly the analysis of qualitative data, or counts, or values on a 1 to 5 scale could use facilities for generalised linear models. We describe the procession of qualitative data in Chapter 18, but the facilities in Instat do not match those in more powerful statistics packages.

In providing the list above we are not trying to undermine Instat. We believe that it is one of the better packages for analysing experimental data, as indicated by the length of this chapter. However, we have seen many, many examples of trials that are over-simplified, mainly because the scientist only has the access or the confidence to use a simple package for the analysis. The key point is that the design should fit the objectives of the study, and the scientist should be confident that an effective analysis is possible for every sensible design. If the analysis is not possible with your usual software then try to change the software, not the design.
Chapter 17 - Regression models

17.1 Introduction

The list of regression dialogues in Instat is shown in Fig. 17.1a.

**Fig. 17.1a The regression menu**

![Regression Menu](image)

We look briefly at the calculation of correlations in Section 17.2. Then, in Sections 17.3 to 17.6 we look in turn at the 4 regression dialogues shown in Fig. 17.1a.

In Chapter 15, we introduced the idea

\[ \text{Data} = \text{Pattern} + \text{Residual} \]

This becomes

\[ \text{Data} = \alpha + \beta \times x + \text{residual} \]

for the example of simple linear regression that we showed in Section 15.5. There our aim was to show how the concepts of simple inference applied equally to regression models. Here, in Section 17.3 and 17.7, we show how an understanding of simple linear regression can also help in teaching many of the ideas that are needed for the more general regression models that we describe in the later sections.

Thus in Section 17.3 the “pattern” consists of a single variate. This contrasts with Chapter 16, where we mainly considered models where the “pattern” consisted of factors. In Section 17.4 we consider regression models where the “pattern” consists of both a factor and a variate.

In Section 17.5 we return to the more traditional regression problems, and consider multiple regression, where there is more than one variate. The final regression dialogue, described in Section 17.6, is where the model includes a mixture of both factors and variates. We look briefly at regression diagnostics in Section 17.7. Then, in the remaining sections, we describe some further topics, such as weighted regression and missing values in regression models.

The last item in the regression menu, shown in Fig. 17.1a, is called “log-linear models”. Readers may be unfamiliar with this method, and surprised to see it with the regression dialogues. We describe this dialogue in Chapter 18. But in the last section of this chapter we consider generalised linear models briefly. This is an important advance, because it allows all that we do in regression models to be generalised to situations where the residual comes from a range on distributions other than the normal.

It turns out that log-linear models are one example of a generalised linear model. This is the only one that is within Instat, so you have to move to more powerful statistical software to
use generalised linear models properly. But this does mean that an understanding of regression modelling is now valuable for two reasons. The models are useful in their own right and they are a useful preparation for understanding the important subject of generalised linear models.

17.2 Correlations

To illustrate correlations and simple linear regression we use the regress.wor dataset, introduced in Chapter 3 of this guide. A plot of the data is shown in the next section, Fig. 17.3b. We see that the yield and day number are highly negatively correlated. The correlation dialogue is shown in Fig. 17.2a, and the results are in Fig. 17.2b. With two columns we included the option to give a confidence interval for the correlation, and also a test of whether the true correlation could be zero.

![Fig. 17.2a Stats ⇒ Regression ⇒ Correlations]

![Fig. 17.2b Results](image)

This dialogue can be used with more than two columns and it then displays the correlation matrix in lower triangular form. It calculates the usual product-moment coefficients of correlation.

An alternative is rank correlations that are sometimes used when the data are not from a normal distribution. This is an example where there is no single dialogue. It is easy to calculate ranks in Instat, using the Manage ⇒ Transformations ⇒ One to One dialogue and one option is to transform the data into ranks. Then use the correlation dialogue, as above, on the ranks. An alternative is to type the commands, as described in Chapter 4, Section 4.5. The data are shown in Fig 17.2c and the commands and results are in Fig 17.2d.
17.3 Simple linear regression

In Chapter 15 we used the same dialogue shown in Fig. 17.3a to consolidate the concepts from that chapter. It also served to introduce the ideas of an Analysis of Variance table. This was in preparation for Chapter 16, but we will see here that ANOVA tables are also useful in regression studies.

![Stats ➔ Regressing ➔ Simple](image)

The graph in Fig. 17.3b includes a plot of the fitted model, together with confidence bands for both the regression line and the points themselves.

The numerical results show the fitted line to be:

Yield = 19.22 - 0.118 * Day

Hence we estimate that the mean yield is reduced by about 0.12 for each day that planting is delayed.

The third type of graph that is possible with the simple regression models is called a "diagnostic plot". We see from Fig. 17.3c that three types of graph are possible, directly from the menu. We choose to look at the residuals against the fitted values and we see, in Fig. 17.3d, that there is a hint that a straight-line model is not adequate. This hint follows from the fact that the residuals seem to alternate, with a block of negative ones, followed by...
positive residuals. This corresponds to the fitted line plot, seen earlier in Fig. 17.3b, where we again see the residuals seem to alternate.

Fig. 17.3c Producing residuals

Fig. 17.3d Graphs of residuals

We will look at another option for plotting the residuals in the next section and the general subject of regression diagnostics in Section 17.7. In Section 17.8 we will try fitting a polynomial model to these data, to see if the fit is improved.

17.4 Adding structure – comparison of regressions

We now return to the survey example that we used in the previous chapters. Our interest with these data is to examine the relationship between the yield of rice and the inputs. In Chapter 15, Section 15.5, we used the simple linear regression dialogue to show there was a relationship between yield and fertiliser, with a fitted model of:

\[ \text{Yield} = 27.66 + 8.948 \times \text{Fert} \]

Then in Chapter 16, Section 16.2, we saw there was also a relationship between yields and the variety of rice used. In that Section we also saw our first “diagnostic plot” which indicated we should consider both fertiliser and variety together. We use the simple regression dialogue again and show the diagnostic plot the “other way round” in Fig. 17.4a and 17.4b. The graph indicates, as expected, that the model with just fertiliser could be improved if variety were also included.

Fig. 17.4a More residuals

Fig. 17.4b Graph of residuals
Hence we move from the simple linear regression dialogue to the Statistics ➔ Regression ➔ Simple with Groups dialogue, as shown in Fig. 17.4c. Here we can include both the Fertiliser and the Variety in the model. We find this dialogue is particularly useful for teaching purposes, because, as we show below, it is an example where we can describe the general concepts of choosing the appropriate model. For the alternatives considered here, the models can all be shown graphically, and that is what the dialogue provides.

Fig. 17.4c  Stats ➔ Regression ➔ Comparison

We investigate 4 possible models and choose the option to plot them. The plots are in Fig. 17.4d to 17.4g. Numerical results are provided to help us choose between these models, but the key information is also given in the title of the plot, if the “Summary” option is chosen in the dialogue shown in Fig. 17.4c.

Fig. 17.4d  Single line  Fig. 17.4e  Common intercept

The first model, in Fig 17.4d, is a single line, for yield against fertiliser, and here it is plotted with a different symbol for each variety. The residual sum of squares from this model is 1961, with 34 degrees of freedom. This is as was found earlier, in Fig. 15.5d and Fig. 15.5h. The 34 degrees of freedom arise because there were 36 observations, and we need 2 degrees of freedom for the intercept and slope of the line.

The next 2 models have a separate line for each variety, but with a restriction. We see that the model with parallel lines, Fig 17.4f, has a lower residual sum of square, of 732, than that with common intercept, Fig 17.4e.
Fitting 3 separate lines, shown in **Fig 17.4g**, hardly drops the residual sum of squares and uses two more degrees of freedom. So we are inclined to accept the model with parallel lines, i.e. **Fig 17.4f**.

Some of the numerical results are shown in **Fig 17.4h**. The first section gives the equation for the parallel line model.

**Fig. 17.4h** Fitted model and reason for the model

| Param. | Estimate | SE  | t    | Prob>|t| |
|--------|----------|-----|------|-----|----|
| Const  | 25.964   | 1.435| 18.08| 0.000|
| Fert   | 5.2643   | 0.9555| 5.51 | 0.000|
| NEW    | 21.791   | 3.042| 7.15 | 0.000|
| OLD    | 9.7227   | 1.987| 4.89 | 0.000|

For factors the estimates are differences relative to the level that is omitted.

Residual S.S. = 718.024 Residual d.f. = 30
Increase in Reg.S.S. = 14.2587 Increase in Reg.d.f. = 2
F-ratio for change = 0.2979 on (2,30) d.f.
Fin> = 0.7445

The equations are as follows

\[
\text{Yield} = 26.0 + 5.26 \times \text{Fert} \quad \text{for traditional rice}
\]

\[
\text{Yield} = (26.0 + 21.8) + 5.26 \times \text{Fert} \quad \text{i.e. 47.8 + 5.26 \times \text{Fert} \ for \ the \ new \ variety}
\]

\[
\text{Yield} = (26.0 + 9.7) + 5.26 \times \text{Fert} \quad \text{i.e. 35.7 + 5.26 \times \text{Fert} \ for \ the \ old \ variety}
\]

In the second part of **Fig. 17.4h** the F probability of 0.74 gives no evidence that a more complicated model is needed. The interpretation of the parallel line model is that the effect of fertiliser is the same for all the varieties.
In the equations in Fig. 17.4h it is slightly awkward that the coefficients for each line are not given directly. If you choose the option in the dialogue, in Fig 17.4c, to give absolute, rather than relative estimates, then the graphs and ANOVA are the same, but the model is fitted without an initial constant. So the output for the model with parallel lines is given directly, as shown in Fig. 17.4i.

**Fig. 17.4i  Results fitting model without intercept**

<table>
<thead>
<tr>
<th>REGRESSION COEFFICIENTS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Y-variate: Yield</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Param.</td>
<td>Estimate</td>
<td>SE</td>
<td>t</td>
<td>Prob&gt;</td>
</tr>
<tr>
<td>Pert</td>
<td>5.2643</td>
<td>0.9555</td>
<td>5.51</td>
<td>0.0000</td>
</tr>
<tr>
<td>NEW</td>
<td>47.755</td>
<td>3.216</td>
<td>14.85</td>
<td>0.0000</td>
</tr>
<tr>
<td>OLD</td>
<td>35.687</td>
<td>2.117</td>
<td>16.86</td>
<td>0.0000</td>
</tr>
<tr>
<td>TRAD</td>
<td>25.964</td>
<td>1.436</td>
<td>18.08</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

This dialogue is mainly to support teaching, because it is limited to models that include a single variate, and a single factor. In practice we usually wish to consider more alternatives. In this example we might wish to also consider the size of the field (another variate) and the village (another factor). We return to this example in Section 17.6.

As a teaching tool we find this dialogue to be useful, because we are able to discuss three key concepts. They are as follows:

- The first is that we need to choose between the alternative models, shown graphically in Fig. 17.4d to Fig. 17.4g. That idea of choice reviews the concept that good models leave as little unexplained variation as possible, i.e. the residual sum of squares should be small. So we are able to show the value of the Analysis of Variance as a tool to help us choose between models.

- In Chapter 15, when we considered simple models, we were not too enthusiastic about significance tests, and preferred confidence intervals. Here, however, we see a clear role for significance testing. The F-test above is a very useful tool to help us in our choice between the different models.

- The analysis should relate to the “structure of the data”. This often means that we should include factors in the regression models. Here is a simple example.

The order of the sections in this chapter is therefore intentional, albeit slightly unusual. Most courses in regression, and the regression section in most statistics packages tempts the move from simple linear regression to multiple regression (our Section 17.5) Then a long time is spent on the concepts involved in fitting a multiple regression equation. But those models are limited to just variates and we often find in real examples that suitable regression models need a mixture of variates and factors.

We return to this issue in the final section of this chapter.

### 17.5 Multiple regression

To illustrate the ideas of multiple regression we use an example from Mead, Curnow and Hasted [1993], pp 197-201 and the data are shown in Fig. 17.5a. The response variable ‘O2’ is a measure of the production of oxygen in samples of water from the River Thames, ‘Chlor’ is the amount of chlorophyll and ‘Light’ the amount of light. The first step would be to plot the response against each of the other variables, and each explanatory variable against the other, but this has been omitted here.
We now explain the logic of fitting multiple regression models in Instat. The top part of the dialogue, in Fig. 17.5b, is all we need for now. We first specify which column is to be the y variable. Here that is X3. Then we consider which terms might be in the model, and we move those columns into the model terms box, in Fig. 17.5b.

Now we fit a variety of models to sort out which is the most appropriate. Initially we need to fit a model and we choose to consider just ‘Chlor’, see Fig. 17.5c. Then we use either the Apply, or the OK buttons. If we use Apply then the dialogue remains. If we use OK then the dialogue symbol on the toolbar can be used to recall the last dialogue.

The F value, in Fig 17.5d indicates that ‘Chlor’ may be useful in the model (compared with the null model).

We then add the term ‘Light’ to the model, as shown in Fig. 17.5e and 17.5f, which shows that ‘Light’ after ‘Chlor’ is needed.
When we now drop the term ‘Chlor’, as shown in Fig 17.5g and 17.5h, we see that it is not so clear that ‘Chlor’ is needed once we have ‘Light’ in the model.

It looks as if a reasonable model is a simple linear regression on ‘Light’ only, and the regression equation could then be shown to be as follows:

\[ y = -1.18 + 0.0106 \times \text{Light} \]

The contribution from ‘Chlor’ is, however, substantial although not significant at the conventional 5% level. We might therefore prefer the model with both variables. So we fit the model again with both variables, and then use the option to give the estimates, as shown in Fig. 17.5i and 17.5j. We see the fitted equation is then:

\[ y = -1.34 + 0.0118 \times \text{Chlor} + 0.0091 \times \text{Light} \]

One would normally proceed with graphical checks of residuals, etc., but this has been omitted here. It is possible from this dialogue, as will be shown in Section 17.7.

What this example has shown is the way the “Add” and “Drop” facilities of this dialogue encourage interactive modelling. We see the same facilities again in the next section, where we look at more general models.

### 17.6 General linear models

We now use the last, and most general of the 4 dialogues for fitting regression models in Instat. This is shown in Fig. 17.6a and is to fit models that include a mixture of variates and factors.

We use the survey data again and first show the dialogue, in Fig. 17.6a, for fitting the same model that we used in Section 17.4. This shows the dialogue, with the facilities for “Fit”,...
“Add” and “Drop”, used as the ordinary multiple regression dialogue, that we saw in Section 17.5.

Fig. 17.6a Statistics ⇒ Regression ⇒ GLM

Fig. 17.6b Adding to the model

There is no longer the limitation to one variate and one factor that we had in Section 17.4. We illustrate in Fig. 17.6b, by adding X3 into the model. The numerical results indicate that this term is not needed. We could also try to add a second factor, namely ‘Village’, into the model. That is also not required.

This dialogue is designed to support teaching and, in particular, to show how factors are included in regression models. In Fig. 17.6c we use the same dialogue as in Fig. 17.6a, but with two of the options ticked. This enables you to look “behind the scenes” to see what Instat does when you fit a regression model containing factors.

Fig. 17.6c How the model “works”

Fig. 17.6d Data

In Fig. 17.6c there is a field called “Calculated Terms”. This contains the variates but has X8-X10, instead of the variety factor. In Fig. 17.6d we show the corresponding columns in the worksheet. Comparing X5 with X8 we see that X8 contains a 1, whenever X5 is “NEW”, and similarly with X9 and X10. These are sometimes called “Indicator variables”.

220 28/06/2002
The factor X5 has 3 levels, and therefore uses an extra 2 degrees of freedom in the regression model, compared with a model that just has an overall mean value for all 3 varieties. The “Fitting Options” in Fig. 17.6c allow us to choose whether to omit the first or the last of the 3 levels. The default is to omit the first level, and so only “X9,X10” are in the “Fitted Terms” field. We could try the option to “Fit all levels” but Instat would then throw out the last one. If we chose the option, in Fig. 17.6c, to not fit a constant, then we could fit all 3 levels. (That is the same as the option, mentioned in Section 17.4, to give the “Absolute”, rather than the “Relative” estimates.)

Once you know the background to fitting regression models by getting indicator variables, it is possible to use the Multiple regression dialogue in the same way. The Manage ⇒ Data ⇒ Indicator Variables dialogue, shown in Fig. 17.6e can be used to give the same columns as in X8-X10 above, or to give further columns that can be used to fit the terms corresponding to the interaction between Fertiliser and Variety, that was also used in Section 17.4. This is done automatically in the General Linear Models dialogue, if you select any interaction terms, and we show that option in the dialogue in Fig. 17.6e. The resulting columns are shown in Fig. 17.6f and show that these interaction columns just multiply the 1’s in the indicator columns by the amount of fertiliser applied.

Fig. 17.6e Manage ⇒ Data ⇒ Indicator

If you use the General Linear Models dialogue, then all these steps are done for you. However the indicator columns still need columns in the worksheet. So you may have to use Manage ⇒ Resize to give the worksheet enough working columns. With the limit of 127 columns there will also be regression problems involving factor columns that can exceed Instat’s capacity.

Most other statistics packages offer the facility to fit a general linear model and they do not suffer from this limitation, because they calculate the indicator columns internally. However, they may still run out or memory or take a long time. Fitting a regression model with 100 terms still requires effectively the capability to invert a 100 by 100 matrix.

17.7 Regression diagnostics

An important feature of interactive regression modelling is the examination of residuals. Basically, residuals are the differences between observed y-values and the corresponding fitted values, usually standardised in some way. Instat can calculate regression residuals in a variety of ways, as we describe below. In Fig. 17.7a we show the dialogue again for simple linear regression, with our usual survey data. Many of the ideas of regression diagnostics can be illustrated with this simple model.
For the usual significance tests and confidence intervals derived from regression calculations to be valid, your examination of residuals should include some check on the normality of their distribution. A popular graphical technique for this is to plot the residuals against their ‘normal scores’. This scatter plot should be close to a straight line if the residuals are approximately normally distributed. This is one of the plotting options of each of the regression and analysis of variance dialogues. The probability plot is shown in Fig. 17.7b, for the regression example fitted using the dialogue in Fig. 17.7a. The assumption of normality seems reasonable.

We next consider the different columns that can be saved that are useful for regression diagnostics. Those that can be saved from the simple linear regression dialogue are shown in Fig. 17a. The results are in Fig. 17.7c, where we used Manage ⇒ View Data ⇒ Format to tidy the display. In Fig 17.7d we show that some extra options can be saved from the multiple regression dialogue.

The remainder of his section is more technical than elsewhere in this guide. The number of observations (cases) is given by \( n \); and \( n=36 \) for the survey data. Then \( p \) denotes the number
of terms in the regression model (explanatory variables) including the constant, so \( p = 2 \) for simple linear regression. \( X \) is the \( n \times p \) matrix consisting of the columns of explanatory variables. If the constant term is in the model, then the first column of \( X \) is a column of 1's.

1. **Leverage values, called LEV in Fig. 17.7c.**

The diagonal elements of the matrix \( X(X^T X)^{-1}X^T \), have special importance. If the \( i^{th} \) diagonal element, denoted \( v(i) \), is large, then the \( i^{th} \) observation may have undue importance in determining the regression model. That is, it exerts a large influence, or a high "leverage" on the model.

The \( v_i \) depend only on the \( x \)-values, not on the \( y \)-value. A large leverage thus indicates an "unusual" combination of \( x \)-values, in the sense that it is different from the bulk of the observations.

A commonly used rough rule for assessing leverage values is to pay attention to cases with \( v(i) > 2p/n \) and to be especially careful with observations with \( v(i) > 3p/n \). The average value of the \( v(i) \) is \( p/n \) and, provided the constant term is in the model, \( v(i) \) is always > 1/n. The absolute maximum possible value of \( v(i) \) is 1.

As an example we consider the leverage values that were saved in X11 in Fig. 17.7c. There is no automatic plot of leverage values but it is easy to use Instat's graphics dialogues. We first use **Manage ⇒ Data ⇒ Regular Sequence** to generate a column from 1 to 36, and then use the **Graphics ⇒ Plot** dialogue, as shown in Fig 17.7e. In this case \( p/n = 2/36 = 0.0556 \) and we see that no value in Fig. 17.7f is greater than twice this value.

![Fig. 17.7e Graphics ⇒ Plot](image)

![Fig. 17.7f Leverage values](image)

2. **Residuals**

a) **Ordinary residuals, called RES in Fig 17.7c.**

The ordinary residuals, defined as

\[
\{ e(i) = \text{observed } y(i) - \text{fitted } y(i) \}
\]
can also be saved from the simple linear regression dialogue and are shown in X9. We see, for example, that for the first observation, the observed value was 53.6, and the fitted value was 50.0, so the residual is 3.6.

b) **Standardised residuals, called SRE in Fig. 17.7c.**

These are defined as the ordinary residuals divided by an estimate of their standard error. They are calculated as \( \{ r(i) = e(i) / (s \sqrt{1-v(i)}) \} \) where \( s^2 \) is the residual mean square and \( v(i) \) are the leverage values, described above. Standardised residuals have standard error...
equal to 1, and as a rough guide, can be considered large if outside the range ±2. Extreme values indicate cases where the model does not provide a good fit to the y-value, or outliers.

The standardised residuals are used in the regression dialogues to give the diagnostic plots. For example, the first option, of a graph “against fitted values”, shown in Fig. 17.7g could equally have been given by plotting X11 against X8.

Fig. 17.7g Residuals against fitted values

<table>
<thead>
<tr>
<th>Standardised residuals against Fitted values</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image.png" alt="Image" /></td>
</tr>
</tbody>
</table>

### c) Studentised residuals

The Studentised residuals are similar to the standardised residuals except that a different estimate of standard error is used. The standard error of $e_i$ is estimated from the data with the $i$th observation removed. Outlying values of the response variable will have large Studentised residuals.

Since the estimate of standard error is independent of $e_i$, Studentised residuals have a t-distribution with $(n-p-1)$ degrees of freedom. (For this reason, if saved, then they are called TREsiduals). This is not true of standardised residuals.

There is some confusion in the naming of types of residuals. What we have called standardised residuals are sometimes called Studentised residuals. Cook and Weisberg [1982] use the terms "internally Studentised" and "externally Studentised" for our standardised and Studentised, respectively.

### d) The PRESS statistic

Yet another kind of residual is defined as follows. The $i$th predicted residual is defined as the ordinary residual obtained from fitting the model to the data with the $i$th case removed. Since the data on the $i$th case is not used in obtaining the $i$th fitted value, the $i$th predicted residual can be thought of as a measure of how well the model can predict the $i$th y-value. A large predicted residual indicates a poor predictive performance at that observation.

The PRESS (predicted residual sum of squares) statistic is defined as the sum of squares of the predicted residuals. It is used mainly in model selection procedures, a smaller PRESS value indicating a better fit. Such a procedure is described in Draper and Smith [1981]. It can be argued (Cook and Weisberg [1982]) that this procedure tends to favour models which fit relatively well at remote or "unusual" observations (those with large leverage).
The display option in the multiple regression dialogues only give the value of PRESS. The actual predicted residuals can be calculated from the ordinary residuals, e(i), and the leverages, v(i), i.e. ith predicted residual = e(i) / (1 - v(i)), i.e. X9 / (1 – X11) in Fig 17.7c above.

3) **The Durbin-Watson statistic**

An important condition for the validity of the estimates obtained in least squares regression is that the residuals are uncorrelated. It is not uncommon, especially in data obtained sequentially in time, for this assumption to be violated. The Durbin-Watson statistic, d, can sometimes be used to test the hypothesis of no serial correlation between the residuals. Critical values of d are given in special tables, see for example Draper and Smith [1981].

4) **Measures of influence**

Leverages give us an indication of "unusual" combinations of explanatory variables, and residuals indicate observations with an unusual response. It is sometimes useful to have an overall measure of how "unusual" an observation is, combining both explanatory and response variables. What is needed is a measure of how influential each observation is in determining the regression. Two options are available for this type of analysis.

a) **Cook's distances, called COO in Fig. 17.7c.**

This statistic combines the standardised residuals and the leverages in a formula which ensure that the result can be compared with an F-distribution. The formula is

Cook's distance  = (1/p) * r(i)^2 * (v(i) / (1-v(i)))

where r(i) is the ith standardised residual, (the column called SRE in Fig. 17.7c). The ith Cook's distance is a measure of the change in the estimated regression coefficients that would result from deleting the ith observation from the analysis. The formula shows that this will be large either if r(i) is large, or if v(i) is large (close to 1). Cook's distances are always positive (they are in fact a kind of squared distance).

An observation with a large Cook's distance can be thought of as worthy of investigation. To decide how large Cook's distance can be, compare their values with the F-distribution on p and (n-p) degrees of freedom. The statistic does not actually follow the F-distribution, but F-values provide a useful scale for detecting influential cases. Cook and Weisberg (1982) suggest that values greater than the 50% point of the F(p, n-p) distribution usually provide a good indication of peculiar observations.

Evaluation of Cook's distance can be recommended as a routine procedure in regression analysis. Instat does not provide an automatic plot, but it is easily given with the **Graphics ⇒ Plot** dialogue, see **Fig. 17.7e** and use X12, rather than X11. If we want first to find the F value, then use **Statistics ⇒ Probability Distributions**, as shown in **Fig. 17.7h**, for the simple linear regression example above. The result is found to be 0.71. The graph of Cook’s distances, in **Fig. 17.7i**, shows that in this case no observation gives cause for concern.
Whereas Cook's distances measure changes in the regression coefficients resulting from deleting each observation in turn, the statistic known as DFFITS measures corresponding changes in the fitted values. DFFITS combines Studentised residuals and leverages by means of the formula,

\[ \text{DFFITS} = t(i) \times \frac{v(i)}{1 - v(i)} \]

where \( t(i) \) is the \( i \)th Studentised residual.

Values of DFFITS greater than \( 4p/n \) are often considered unusual. Unlike Cook's distances, DFFITS can be positive or negative.

### 17.8 Lack of Fit

In Chapter 12 we looked at the fertiliser levels used by the 36 farmers in the rice survey. This gave the table shown again in Fig. 17.8a, showing that only integer, or half integer values were used.
So instead of fitting a simple linear regression, we could derive a column from X4, using Manage ⇒ Calculate, as shown in Fig. 17.8c. This could then be used as a factor, see Fig. 17.8d.

Then we use the one-way ANOVA, from Chapter 16, to look at the effect of fertiliser on yield. The dialogue is shown in Fig. 17.8e, with a graph of the means in Fig. 17.8f.
It is this graph that we fitted a straight line to, in Chapter 15. Now we can use the repeated values at the different fertiliser levels, shown in Fig 17.8f to see if the straight line is adequate.

We use the Statistics ⇒ Regression ⇒ Simple dialogue again, but this time with the option for “Lack of fit” as shown in Fig. 17.8f. We see that the residual mean square and degrees of freedom from the last ANOVA have been transferred to this dialogue. If it is not the last ANOVA that is used, then you have to enter the appropriate values.

The results are given in Fig. 17.8 and do not imply a lack of fit.

17.9 Polynomial and periodic regression models

Polynomial regression models can be fitted by regarding them as a special case of multiple regression. If a plot of the response variable, \( y \), against an explanatory variable, \( x \), suggests that a polynomial regression might fit the data, you have to create new variables representing the powers of \( x \) required, and then use them as ordinary variables in the multiple regression dialogue.
As an example we consider the simple regression example from Chapter 3 that was used earlier in Sections 17.2 and 17.3. There we noted that a straight-line model might not be adequate. We can use the Manage ⇒ Calculate dialogue to give the polynomial terms, but the Manage ⇒ Data ⇒ Polynomials dialogue, shown in Fig. 17.9a, is easier. We choose to use simple polynomials. This gives the worksheet shown in Fig. 17.9b.

One potential problem with the simple polynomials is that the successive terms are highly correlated. We can see this by using the Statistics ⇒ Regression ⇒ Correlate dialogue, shown in Fig. 17.9c. The results, in Fig. 17.9d show the correlations between the polynomial terms are from 0.90 and 0.98. The regression algorithm can cope with this, but we would not want to go to a higher power without perhaps using the orthogonal, or at least the centred polynomials.

We now use the multiple regression dialogue. We begin, as shown in Fig. 17.9e, by fitting a straight line and then "Add" the quadratic and cubic terms.
The results are shown in Fig. 17.9f and indicate that a cubic model does fit better than a straight line. We therefore use the option in the dialogue to give the estimated coefficients of the cubic equation.

**Fig. 17.9f Results**

From Fig. 17.9f we see that the cubic equation is

\[
Yield = 17.258 + 0.067 \cdot \text{day} - 0.0032 \cdot \text{day}^2 + 0.00001 \cdot \text{day}^3
\]

This is an example where Instat's default printout is inadequate. We have tried to avoid scientific notation where possible, because it can be confusing in an introductory package. Here, however there is only 1 significant figure for the cubic term. We therefore type the command

: ACCuracy 5

and get the estimates again. We can now go back to the dialogue, or just type

: ESTimate

as another command. The result is displayed as

\[
Yield = 17.258 + 6.7 \cdot 10^{-2} \cdot \text{day} - 3.2048 \cdot 10^{-3} \cdot \text{day}^2 + 1.4624 \cdot 10^{-5} \cdot \text{day}^3
\]
If you are allergic to giving commands, then an alternative is to use the save option in the multiple regression dialogue and save the estimates into a column. They are saved with more precision than they are displayed.

Once a polynomial equation has been fitted to the data, the save option can be used to save the fitted values. They can then be plotted, together with the observed data, as shown in Fig. 17.9g, with the resulting graph in Fig. 17.9h. Alternatively, the equation can be typed into a string, e.g. S1, in the worksheet, and plotted, together with the data, as we showed in Chapter 11.

It is possible to fit other types of equation to data by means of similar techniques. For instance, trigonometric functions of an x-variable may be useful in fitting equations to periodic data. In this case the dialogue Manage ⇒ Transformations ⇒ One to One, shown in Fig. 17.9h, is an alternative to the Manage ⇒ Calculations dialogue to produce the x variables.
17.10 Weights and missing values in regression

The multiple regression dialogue allows a column of weights to be specified. Then all subsequent regression models will be fitted by weighted least squares, with weights in Xm. Weights must be non-negative.

Zero weights are allowed, and their effect is to delete zero-weighted cases from the regression. This is a useful way of dealing with outliers without having to alter the data in any permanent way. Total and residual degrees of freedom are adjusted for zero-weighted cases. It is also possible to recode cases that are to be removed into missing values. The effect on the regression is the same as zero-weighting.

The CORrelation dialogue can use zero-one weights as described above, in order to remove cases from the calculations. Note that, in this case, the weights can only be zero or one. Otherwise weighted correlations are not allowed.

If there are any missing values in the potential columns specified in a regression model, then this is taken into account in all subsequent regressions.

Thus if there is a missing value in a particular row in any one of the variables (or in the weights column, if there is one), then that entire row is omitted from all subsequent regression models. This remains true even if the particular column with missing values is not included in the particular regression model. For example, suppose, in Fig. 17.10a, there is a missing value in row 8 of column X1. Then row 8 will not be included in the regression calculations, even though X1 is not in the current model.

Finally, we consider the fitted values and residuals when there are missing values. If there are missing values in any of the x-variables (or in the weights), then the corresponding fitted values and residuals are set to missing values also. However, if there are missing values only in the y-variate, then the fitted value is calculated, though the residual is set to missing.

17.11 In conclusion

As with the last chapter, we consider the regression facilities in Instat to be one of its strengths. This is not only for teaching purposes, but also for serious data analysis. In particular we emphasise the value of being able to consider regression models where there is a mixture of variates and factors in the model.

We now consider some of the topics that would be available in more powerful packages. The first is that of automatic selection of variables. Regression courses, and statistics packages emphasise a range of methods to help the computer to choose the regression model automatically. This version of Instat just provides “Add” and “Drop” and expects you to do the work.

This lack of “stepwise”, and “all-subsets” regression is not a great lack for the number of variables that are usually considered when Instat is used. But, if you like automatic methods, then check that they are provided for the types of model that you wish to fit. Some
packages provide a plethora of methods for multiple regression models where there are no factors. Then separately they have a GLM (General Linear Models) dialogue, but without any support for variable selection. Really nothing. Compared to Instat they just provide "Fit", and not even "Add" and "Drop".

In our view this does not encourage good practice, because large problems (of the type where automatic methods might help) are likely to include some structure and hence might need factors in the model.

When a straight line might not be sufficient, we have shown in Section 17.9 that it is easy to consider polynomial models. There are now many alternatives to polynomial models and some packages, like S-Plus, Genstat, Stata and SAS also include spline models, and other robust methods, such as Loess fitting.

The regression menu from Genstat is given in Fig. 17.11a, and shows how much more is offered by some packages. They have 9 options under their regression modelling, and everything in this chapter is within their first option, called "Linear". This option includes stepwise fitting for models that can have with factors, models that include splines and also an important "Predict" dialogue.

The other options also typify further models that can now be fitted routinely. We show the dialogue for the second option in Fig. 17.11b. Instat's regression modelling is limited to situations where the residuals can be assumed to be from a normal distribution. If not then the common "old-fashioned" solutions were either to transform the data, so the residuals were roughly normal, or to abandon the modelling approach and move to non-parametric methods. We look at non-parametric methods in Chapter 18.

Now, as indicated in Fig. 17.11b, a data from a range of other distributions can be modelled by regression methods. They are called generalised linear models and have been around for about 30 years. Some packages, like Genstat or Stata consider them as an extension to regression modelling, and hence all the ideas of this chapter apply. This includes the possible need for factors in the explanatory variables, automatic variable selection and so on.

Other packages treat particular cases separately. For example logistic regression modelling is for the situation where the y-variable takes just 2 values, "success" or "failure", or 0 and 1.

Either way, this is a major advance and one that has become within the reach of more non-statisticians. As we are now considering more advanced methods, it is not surprising to find that different packages differ in their approach. The different approaches also follow partly
because many of the major statistics packages were already in existence before these methods were invented.

What we find exciting is that the ideas of regression modelling, that we have described in this chapter, now open the door to such a wide class of models. They should be used more than they are.

There are also options for "non-linear regression modelling" in the Genstat menu in Fig. 17.11a. By "non-linear" we mean a fitted model that is more complicated than \( y = a + bx + cz \). So a quadratic curve counts as a "linear model", because we can define \( z = x^2 \).

Examples of non-linear models are

\[
Y = \frac{a + bx}{c + dx}, \quad \text{or} \quad y = a\exp(bx).
\]

Again the standard packages differ in what they offer. In particular many packages seem to feel that when non-linear models are fitted, that the data does not have any structure, so factors are not needed as part of the model. In our view this is rarely the case. So, if you need non-linear models then check also how your chosen package copes with factors in the model.
Chapter 18 - Qualitative data

18.1 Introduction

Researchers often collect "qualitative data". For example a research project was interested in looking at how farmers use water. In a development context, this study would probably include initial discussions with farmers using participatory methods, focus groups or semi-structured interviews. The qualitative data from this process might show that farmers use water for irrigation, animals, household, washing clothes, transport, generation of electricity, milling, festivities, and so on.

If the information is just the list of different uses that one farmer, Mr Tambo, makes of water, then statistics offers little help in analysing the information. The information may simply be that Mr. Tambo uses the water for irrigation, for his household and to wash clothes. However, statistics may help if a larger number of farmers have been asked the same questions and the answers have been recorded appropriately. There were 46 farmers in the example used later in this chapter.

Often a list of uses, such as is given above, is not the only result that is needed from the research. Let us assume that those 46 farmers are distributed in a region with 3 well-defined agro-ecosystems, and have different educational levels. The researchers would then often wish to look further at these data. Perhaps they might be interested in selecting say the four main uses of water and going back to the farmers to get an idea of the level of importance that they would attach to different uses and why. The type of information that would come from such a research project and some of the analysis that can be done is the topic of this chapter.

We use statistical ideas when we need to summarise the information that comes from these farmers. One way of summarising this information is by counting how many times a particular water use has been quoted by farmers. These are frequencies and we can analyse frequencies. In addition, if farmers have been asked to rank four uses in order of importance, the results of this ranking can be analysed using statistical techniques, as we illustrate in this chapter.

So, statistics helps in dealing with qualitative data when the amount of data is large and we need to summarise the information in order to proceed with the analysis. This process usually involves counting or grouping and yields numerical quantities that can be dealt with using statistical techniques.

For an analysis to be useful there are some basic requirements in terms of how the information are collected. These are the same that we would assume have been fulfilled when any information is collected for research purposes. In particular we assume that the information on different subjects (individuals, communities, plants, animals, etc) is of the same nature and has been collected using a consistent methodology.

We also assume that the subjects have been selected in such a way that they are a good representation of the population that is of interest for the researcher. We also assume that the information has been managed properly from the source to the records. This means that there has not been loss of information due to transcription, storage or human intervention.

In its raw form the qualitative data may be available as texts, maps, tape recordings, or in a variety of other forms. An initial stage in processing these data is to code the information. This is only worth doing if there is a reasonable volume of data, otherwise the results are simply reported individually. Coding will lose information, and therefore it must be done carefully.

Once the information from any subject has been coded, it is then important to examine the raw record again, to see what information the coding process has not captured. What is left
over might be called the "qualitative residual". Sometimes this examination will lead to more codes. Otherwise it may lead to individual reports of particularly interesting responses. These reports are in addition to the processing of the summary data. This reporting of either the left-over information, i.e. the "qualitative residuals", or of the "odd" or "individual" responses themselves is the same as the requirement to report interesting or odd "outliers" in the analysis of ordinary (quantitative) data.

The codes are sometimes called categories, and the subsequent analysis is then the "analysis of categorical data". Often the initial step will result in more than one code per response. In this example, it might be a list of the different uses of water. In the chapter on tables we looked at how to process multiple responses in Sections 13.8 and 13.9.

18.2 Categorical data

In some studies part of the data are collected initially as categories. Here are some examples:

Attendance at lectures should be compulsory

| Agree | Disagree |

Please rate this lecture on the following scale

| Very good | Good | Average | Bad | Very bad |

With the pictures as standards, score these gardens on the scale 1 to 9.

| Score |

Please give your age in the following categories

| 17 or less | 18 to 29 | 30 to 54 | 55 or over |

Please state your main occupation

Please rank the following 4 uses of water in order for you (1 is the most important)

| Irrigation | Livestock | Household use | Clothes Washing |

Using anywhere between 0 and 10 buttons, please indicate how important the following uses of water are to you.

| Irrigation | Livestock | Household use | Clothes Washing |

We have seen other examples in this guide already. In the rice survey, used in many chapters, X5 is the variety of rice used by the farmers, and is coded between 1 and 3. These are often called "categorical data". In Instat these variables are usually made into factors, and can then have descriptive labels attached. Sometimes, as we have seen, these
factors define the structure of our data, and we analyse other variables separately for each level of these factors. In this chapter these are the columns to be analysed. They are now “centre-stage” rather than being in a supporting role.

Often these data are ordered and so are called “ordered categorical data”. For example, the question on the quality of a lecture is answered on a 1 to 5 scale, with 5 corresponding to “Very Good”.

The reason the analysis is different is that we do not want to use arithmetic, i.e. to say that 4 = 2 + 2. In this example that would correspond to the idea that a rating of 4, i.e. “good” being twice a rating of 2. If we do not want to use simple arithmetic, then we do not want to take means and variances, because they involve adding the values.

If the data are ordered, then we are happy with idea that 4 is larger than 2. From there it follows that some of the methods we will use in this chapter will involve ranking the data. Sometimes the data we collect are ranked from the outset, because we have asked respondents to put items in order of importance. One of the examples above was of that type.

In Section 18.2 we consider how to teach this subject and we introduce some of the examples that will be used in the remainder of this chapter. We suggest that the way this subject has been taught in the past has contributed to poor analyses.

Then we look at different methods of processing the data. In Section 18.3 we look at the dialogues in Instat for “non-parametric”, or “distribution-free” analyses. The menu is shown in Fig 18.2a.

Fig. 18.2a  Non-parametric menu

In Section 18.4 we examine the chi-square test menu and extend these ideas to log-linear modelling in Section 18.5. The dialogue for log-linear models is under Statistics ⇒ Regression. This sets the scene for a discussion of further approaches that are in the final section of this chapter.

Further information on the collection and analysis of some types of qualitative data is in our good-practice guide called:

“Analysis approaches in participatory work involving ranks or scores”
18.3 Examples and teaching

**Fig. 18.3a** shows the rice survey data again to review some ways of processing qualitative data that we have already seen. In previous chapters we have mainly been concerned with the analysis of the yields, but here we look more at the qualitative information, of which the variety column in **Fig. 18.3a** is an example. In this figure we also see that quantitative columns, like the fertiliser level, given in X4, can be recoded into ordered categories, and an example is shown in X7.

In **Fig. 18.3b** we show a simple frequency table of the variety of rice used by the 36 farmers. This can be found using the methods described in Chapter 13, on tabulation, and Chapter 12, on data summary. We used the `Statistics ⇒ Tables ⇒ General` dialogue to give **Fig. 18.3b**.

In Chapter 15 we considered simple inference, with the menu shown in **Fig. 18.3c**. There we showed how the binomial model could be used for answering questions concerning the variety data.

**Fig. 18.3a** Survey data

**Fig. 18.3b** Varieties of rice

**Fig. 18.3c** Menu for proportions

**Fig. 18.3d** Proportions –raw data
We look here at the example of the proportion of farmers using the traditional variety. We can use either the raw data, as in Fig 18.3d, or the summary values, as in Fig 18.3e. We see, in Fig. 18.3f, that the proportion is 0.42 or 42%, together with confidence limits for this proportion.

So we can process qualitative data, using the descriptive methods, described in Chapters 12 and 13. We can also fit simple models using the methods shown in Chapter 15. That is a good start.

In Chapters 16 and 17 of this guide we moved to a range of more complex models, in the chapters on ANOVA and regression. There we made no mention, except in the last sections, of models for non-normal data.

This is the equivalent of many courses and books on statistics. In the early parts of a typical statistics course we discuss different types of data. Then we look at descriptive methods. This is followed by ideas of probability modelling (Chapter 14) and simple inference (Chapter 15). So far we still include qualitative data. However once we move to realistic modelling, we forget all those distributions and just consider the normal distribution.

This description and the figures above provide the background to concerns we have about the teaching and the knowledge that users have, on the processing of qualitative data. There are three main concerns.

Our first concern is the presentation of the examples of data in training courses. In the example above, the raw data are shown in Fig. 18.3a and a summary of the variety column is in Fig. 18.3b. The analysis, with the dialogues in Fig. 18.3d and 18.3e, can either use the raw data as in Fig. 18.3d, or just the summary as in Fig. 18.3e. We strongly suggest that students should start with the raw data wherever possible, and not just with the summaries.

Our concern is that courses often start with the summary values, and rarely look back at the raw data. This was perhaps excusable before computers were used in training courses, but now there is no reason for hiding the raw data.

A good reason for showing the raw data on a routine basis is that the main objectives of most studies cannot be answered from just the summary data. We have the analysis above, that 42% of farmers use traditional rice, but suggest that this will rarely be all we need to know. For example:

- Are those farmers from all the villages, or are some villages more traditional than others?
• Are those farmers of a particular type? For example do they tend to have the smaller fields, or apply less fertiliser. In a real survey we might also ask about their cultural practices in relation to questions concerning education, head of the household, and so on.

So, starting with the summary data may help us to teach particular techniques, but we must be sure that trainees see the full picture as well. And starting by showing the structure of the raw data will help.

Our second concern is the overuse of significance tests, both in training courses and in data analysis generally.

We use a second example to illustrate this point. It is also in the “ranks and scores” guide mentioned in Section 18.1. It concerns a survey on 46 farmers. Within this study they were asked to rank 4 uses of water (irrigation, livestock, household, clothes washing) in order of importance. Part of the raw data is shown in Fig. 18.3g and we see that most farmers rate irrigation to be the most important use of water, i.e. it often has rank 1.

![Fig. 18.3g Data on ranks](image)

![Fig. 18.3h Statistics ⇒ Summary ⇒ Group](image)

The Statistics ⇒ Summary ⇒ Groups dialogue can be used, as shown in Fig. 18.3h, to give a simple summary of the data. The summary values have been saved into further columns in the worksheet, as shown in Fig. 18.3i. Some of the results are also in the Output window, as shown in Fig. 18.3j.
For example, we see from either Fig. 18.3i or 18.3j that 34, i.e. 74% of the farmers rank irrigation as the most important use of water.

In the next section we will use a distribution-free analysis to show that the differences in the ranks could not be due to chance, and this confirms that the perceived importance of irrigation to the farmers in this survey is a real feature of the data.

Our concern is that we should not stop there, because significance tests rarely relate directly to the full objectives of a research study. If we only do significance tests we sometimes run the risk of only proving the “bleedin’ obvious”, to quote the words of Basil Fawlty in the celebrated UK Television series called Fawlty Towers.

In the above context, there is nothing wrong with this test, but it may only confirm what we largely knew before. What might relate more closely to the objectives of the study could be to look at the second choices of those 74% who put irrigation first, and also look in more detail at the other information of those 12 farmers who responded differently.

In the context of this guide we introduced significance testing in Chapter 15, largely to teach the key concepts. Then, in Chapters 16 and 17, significance testing was important, but usually at the early stages in the analysis. It helped us to choose between competing models for the data. Then the analysis proceeds with an interpretation of the parameters of the model that we choose.

The situation is no different here.

And that brings us to our third concern. This is that the processing of qualitative data is taught as a different subject to the analysis of “ordinary” variables. This is sometimes taken to the extent that people collecting qualitative data argue that they do not therefore need “statistics”. We hope that users will see otherwise, by the end of this chapter.

18.4 Non-parametric methods

The median is a key statistic in the methods described here. We have already used the median in data exploration when looking at boxplots and stem and leaf plots in Chapter 11.

The simplest analysis is provided by the sign test. We illustrate with the irrigation data, shown in Fig. 18.3g. Here the data are the ranks, between 1 and 4 that are allocated to each of the 4 columns, with irrigation being the first column. If there is no difference between the importance of the uses of water, then the median rank for irrigation will be 2.5, and this gives a convenient null hypothesis.
Fig. 18.4a  Stats ⇒ Non-parametric ⇒ One and Two sample

The dialogue is shown in Fig. 18.4a, and the results are in Fig. 18.4b. We see that 42 of the 46 farmers gave irrigation a rank less than 2.5, i.e. a rank of 1 or 2, and hence there is no chance that the true median is 2.5.

Fig. 18.4c  Stats ⇒ Simple models ⇒ Proportion, One Sample

An equivalent way of doing this test uses the Statistics ⇒ Simple Models ⇒ Proportion, One Sample as is shown in Fig. 18.4c, with the results in Fig. 18.4d.

To show the way the analysis might continue, we now select just those 34 farmers who stated that irrigation was the most important, i.e. they gave it a rank of 1. We use Manage ⇒ One to One ⇒ Select, as shown in Fig. 18.4e. Then the ranks of the remaining uses of water are 2, 3 or 4, so we see whether the median rank for livestock is 3. The results from the sign test (not shown) are that 16 farmers gave livestock a rank of two and only two farmers gave it a rank of four. Clearly the median rank is not three.
The sign test is simple and intuitive, but for data where there are many categories it only considers whether a value is less than the hypothesised median, or not. The Wilcoxon and Mann-Whitney tests provide non-parametric equivalents to the one and two sample t-tests considered in Chapter 15. We illustrate with the rice survey data, and consider the 2-sample case. We ask whether the fertiliser use is the same for 2 groups of farmers. The dialogue is shown in Fig. 18.4g with the results in Fig. 18.4h. The z value is statistically significant, and the comparison of the medians shows that the farmers using the old improved varieties tend to apply more fertiliser than those using the traditional ones.

There is also a non-parametric equivalent to the one-way ANOVA, called the Kruskal-Wallis test, and the Friedman test for the 2-way ANOVA. We illustrate the latter on the data shown earlier, where the ranks were given for the use of water.

For this analysis the data must first be stacked into a single column, and this uses the Manage ⇒ Transformations ⇒ Stack dialogue, shown in Fig. 18.4i, with the resulting data in Fig. 18.4j.
The non-parametric ANOVA dialogue is shown in Fig. 18.4k and the results are in Fig. 18.4l. We see the large difference in the median ranks, and also the rank sum. Irrigation is clearly the use of water with the lowest rank sum.

**18.5 Chi-square tests**

We use an example from Mead, Curnow and Hasted (1993), page 282. The data are shown in Fig. 18.5a. They have already been summarised to give the frequencies. There the data in X4-X6 are shown as a four by two table, as in the textbook. They are also shown in "column" format, in X1-X3.
There are 4 treatments in this fertiliser study and what has been recorded is the incidence of blackleg (a bacterium) on potato seedlings. In Fig. 18.5b we show the 2-way table that can be used to display the data, if entered in column format. This shows that 44 of the 450 seedlings were infected with blackleg.

In Fig. 18.5c we use the Statistics ⇒ Tables ⇒ General dialogue to present the information in percentage form as might be given in a report. We see, from Fig. 18.5d, that about 16% of the seedlings have blackleg when there is no fertiliser, compared to under 4% for one of the levels of fertiliser.

So far this is descriptive statistics and we have used the tabulation facilities, described in Chapter 13, to give the results. As an inference question we see that this is one where 4 proportions are to be compared, because there are 4 treatments. We have looked at 2 proportions in Chapter 15, so this extends the methods from Chapter 15, with frequency data, just as Chapter 16, extended the one and two-sample tests for normally distributed data.

This is sometimes called a 2 * 4 contingency table and we use the chi-square test to see if the treatments are really different. Either layout of the data described above may be used, as is indicated in Fig. 18.5e. The results are given in Fig. 18.5f. The significance level of
the chi-square test is about 3%, indicating that the observed differences in infestation are probably real.

How are we now to present the results? If they had not been significant it would have indicated that no treatment difference could be detected. So we could have presented just the one-way table of overall infection. Then we could simply report in the text that 44 of the 450 plants, i.e. 9.8% were infected and that the treatments did not seem to have any effect. As the results were significant we have our “passport” to interpret the data on the 4 treatments, using the table shown in Fig. 18.5d.

**Fig. 18.5e Statistics ⇒ Simple Models ⇒ Chi-square**

**Fig. 18.5f Results**

This subject of examining frequency data is sufficiently important that there is a special guide, as part of Instat’s help. In that guide the first example is of a 3 * 4 table, to illustrate a more general contingency table. The data are on hair and eye colour for 6800 males in Baden, and is taken from Mead, Curnow and Hasted (1993). The data are shown in Fig. 18.5g with the results of the analysis are in Fig. 18.54h. The analysis follows from the same dialogue used earlier in Fig 18.5e.
In **Fig. 18.5h** the large value of the chi square shows that there is a clear relationship between hair and eye colour. The analysis should then continue, and would usually include tables of percentages.

In Chapter 13 we saw that we could produce 2-way tables, as in **Fig 18.5g**, but we also saw that 3-way and 4-way tables are also often useful to present and summarise data. We have seen here that the chi-square test can help us to decide on the presentation of one, or two-way tables. In the next section we look at a more general system for analysing frequency data that is not limited to 2-way tables.

### 18.6 Log-linear models

There is a parallel between the dialogue used here and those used in Chapter 16, where we processed “ordinary” data. There the simplest dialogue was for one-way ANOVA. That allows for just a single treatment factor and is the equivalent of the chi-square test described in Section 18.4. Log-linear models provide the equivalent of a general ANOVA system, in that we can process frequency data where there are any number of factors.

This is an important advance for those who are only familiar with chi-square tests.

When there are two factors the Statistics ⇒ Regression ⇒ Log-linear models dialogue is as easy to use as the dialogue for the chi-square test, shown in the last section. We illustrate in **Fig. 18.6a** with the same 4*2 table that was used at the start of the previous section.
The one difference is that the layout of the data must be in the "column" format, i.e. with the data in a single column, and two more columns with the factors.

We now consider a table with three factors and use the example from Mead, Curnow and Hasted (1993), page 334. The data are shown in **Fig. 18.6c**. They are on the frequencies of occurrence of different numbers of lambs per birth, for ewes from three breeds at each of three farms.

The Statistics ⇒ Tables ⇒ Summary dialogue, shown in **Fig. 18.6d**, may be used to show the data in tabular form. This display is in **Fig. 18.6e** and is how they are presented in the textbook.
We show the routine for the analysis and then interpret the results. The first step is to calculate the interaction columns, between the three factors, X1, X2 and X3. We use the Manage ⇒ Data ⇒ Interactions dialogue as shown in Fig. 18.6f, and this results in the extra columns as shown in Fig. 18.6g.

Now we use the Log-linear models dialogue to fit the alternative models. One example of the dialogue is shown in Fig. 18.6h, and the results for three of the models are in Fig. 18.6i.
In this case our objective is to understand the relationships between X3, the pattern of births and the two explanatory factors, X1 and X2. As with the analysis of 2-dimensional contingency tables, our hypotheses usually (as here) mean that we do not question the counts in the margin of the table shown in Fig. 18.6e. So, for example, we do not ask whether the farms have an equal number of births.

So in model terms we consider 4 alternative models, with the simplest being

\[
\text{Frequencies} = \text{Farm} + \text{Breed} + \text{Farm by Breed} + \text{Number}
\]

That would correspond to the model with the factors X1-X3 and X5. This is not shown, but did not give a good fit.

The first results in Fig. 18.6i show that X1-X3 plus X5-X7 is a good fit. This is because the residual deviance is 14.6 with 12 d.f. and that shows we do not need the 3-way interaction term in the model.

The other results in Fig. 18.6i show that we can omit X7, the Number by Breed interaction from the models, but we cannot omit X6, the Number by Farm interaction.

Our interpretation is very similar to the processing of data that is normally distributed, as described in Chapter 16. We use these significance tests to indicate which tables of results to present. In Chapter 16, we usually presented tables of means, while here we often present tables of percentages.

So here the dialogue to produce the resulting table of percentages is in Fig. 18.6j. The table that corresponds to the simplest acceptable model is in Fig. 18.6k and this now has to be interpreted. We see there, that the main cause of the interaction is that 14% of the sheep have triple births in farm 1, and this is far higher than the other two farms. And Farm 3 has a higher percentage of single births.
There is much more that can be said on the fitting of log-linear models, not least why they are called "log-linear". More details are in the guide titled “Fitting of Log-Linear Models” that is part of the Help with Instat. There we also discuss the problem of zeros in the data and consider how a logistic model can be fitted using log-linear modelling.

18.7 In conclusion

In Section 18.2 we concluded with the hope that readers would see the unity of the ideas for data processing by the time they reached this section. Perhaps the best example was the 3-factor contingency table, described in Section 18.5. Here there were effectively 2 treatment factors (the third factor was the response) and the concepts of the analysis parallel those in Chapter 16, where we have also looked at studies with 2 or more “treatment” factors.

Log-linear models are an example of “generalised linear modelling”. For example logistic models are for the situation where the response has just two alternatives and permits the same range of models to be fitted that was described in Chapter 17 on regression methods. Adding facilities for logistic models is high on our wish list for Instat. Currently you can fit logistic models as a special case of log-linear modelling, but that is only when the model just has factors. In general the fitted model could have any combination of factors and variates as we describe for normally distributed data in Chapter 17.

All the major statistics packages have facilities to fit logistic models, but they differ in precisely what they offer. Some, like SPSS and Genstat, offer the same range of variable-selection methods as are available for ordinary regression models.
Chapter 19 - Handling time

19.1 Introduction

The main dialogues we use in this chapter are shown in Fig. 19.1a to 19.1c.

In many studies the important measurements concern time. One possibility is to record the time at which a particular event occurs. These are often called "survival" studies. They are particularly important in medical research, where the event may indeed be the (non) survival of the patient.

But the subject is general. Perhaps the event might be the time to the (non) survival of a packet of cigarettes after an attempt by smokers to give up smoking, or the time to the use of a telephone after a child has been told to start her homework, or the time to germination of seeds after planting.

The alternative is to record events at specific times. Unsurprisingly this is called "time series", see Fig. 19.1b. Many time-series, are to monitor the current state at regular intervals. Examples of monitoring include unemployment numbers, or sales of chocolate, or the state of the ozone layer each month.

Sometimes we may have both types of measurement in our study. For example, we record the weight and health status of people each month and then also note the time until a particular event occurs. But survival studies assume a particular intervention that marks the start of the study. In ecology this might be the date of an oil-spill or the start of a campaign to promote more recycling.

In this Chapter we look at survival studies in Section 19.2 and the analysis of time series in Section 19.3.

There are also many studies where recordings are made in time, but the time aspect is of secondary importance, in the objectives of the study. For example, we may record the weights or milk-yield of animals at regular intervals, to assess the effectiveness of different diets. Or we record the reaction times of drivers at regular intervals after starting them with different quantities of alcohol.

In this type of situation we have many (usually short) time series, with one for each animal or person. These are conventionally called "repeated measure" studies and we consider how such data may be processed in Section 19.4.

Time does not just march on; it keeps coming round. Whether on a daily, or a yearly basis, it is sometimes useful to think of our measurements on a “circle of time”. Often the “time of day” or “seasonality in the year” are a nuisance, and are not related to our objectives. But
sometimes the time in the day when an event occurred is key, for example when in the day do rainfalls start, or when should a car journey begin. There are some special points to be made about the processing of circular data that we consider in Section 19.5.

**Fig. 19.1c Circular statistics**

Both students and teachers of statistics may be surprised to see these four topics in an introductory guide to a simple statistics package. In the literature there are books that are on each of the topics in this chapter. However, service courses in statistics omit all these topics and they would rarely include more than one of them. This chapter is partly designed to question that approach to our teaching and for two reasons. The first is that with access to the computer and statistical software, it is easy to give users a taste of these subjects. The second is that we believe that simple methods are useful for the analyses described here. These are methods that are within the range of Instat, and hence of simple teaching.

We will see that the simple approaches consist largely of being able to organise the data appropriately, as described in Chapters 8 and 9, and then being able to do constructive descriptive analyses, as described in Chapters 11 to 13. Even if the methods described here are not all you need, they are a useful start in any study. Where more is needed, they would provide a good base for collaboration with a statistician.

Some readers may feel that all these types of study sound similar and so you might have a situation where any of these methods might apply. If so, then we have two points that might be comforting. The first is that how you analyse the data depends on your objectives. So, when you look at each method, you can always give it the “would that help my objectives?” test.

The second point is that there is now a more unified approach to the analysis. If data analysis for you still consists of “which significance test should I do?”, then life is becoming confusing. But if you throw away part of your “significance-test security blanket”, then both the simple methods of analysis and the fitting of an appropriate model is done similarly for all these situations. We return to this point in the final section of this chapter. We will see, by then, that the simple methods could be done with Instat, or even with a spreadsheet. A more powerful statistics package is needed for most of the modelling.

### 19.2 The time to events

In this section we consider studies where a measurement of interest is the time to an event. Common examples are in medical studies, where the event may be the recurrence of a condition or the death of a patient, but the subject is general. It could also be the time to sighting of an animal in an ecological study, or the wait for a hospital bed, or the time to breaking of rivets in an engineering experiment and so on.

The data for a simple example, from Collett (1994) are shown in **Fig. 19.2a**. This gives the survival times of two groups of women with breast cancer. The groups were divided according to their status in relation to a marker that might differentiate between two stages of the cancer.
In **Fig. 19.2a** the data are in days, and those observations with a *, signify women who were still alive when the study ended. This is called censored data, because we know the women survived longer, but not exactly how long. The data in Instat are shown in **Fig. 19.2b**. Instat uses numeric codes to represent censoring: 0 indicates no event [patient still alive] and 1 indicate that the event has happened [patient died].

This aspect of censoring is one reason why these data are considered as a special topic. It does not happen in all studies. For example, in a seed germination trial, we can plant all seeds on the same day and consider that any that did not germinate by a certain date are never going to do so. In such cases we can split the data into those that germinated and those that failed. This could involve a binomial model, as described in Chapter 15. Then we can study the time to germination of those seeds that succeeded. This gives a survival time problem, but without censoring.

As in any other study, there are often three stages to the analysis.

The first is a descriptive summary of the data, using the methods described in Chapters 11 to 13. Some special graphs may also be useful, as we describe here, using the Statistics ⇒ Survival ⇒ Kaplan Meier dialogue.

The second stage, in some studies, is to move to simple one and two-sample inference as we described in Chapter 15 for normally distributed data, and Chapter 18 for distribution-free methods. We describe the equivalent analysis below.

The third stage is to fit more general models, as described in Chapters 16 and 17 for normally distributed data.

As an example we analyse the dataset in **Fig 19.2a** and **19.2b**. We use the Statistics Survival ⇒ Logrank, as shown in **Fig 19.2c** and **19.2d**.
The results include the survival curve, shown in Fig. 19.2e. The survivor curve equals one before the first event, as all individuals are alive when they enter the study. It is also assumed to be constant between each event, so the graph in Fig. 19.2e is a step function that decreases immediately after each event.

The survivor function for the negatively stained group always lies above that for the positively stained group, so the two groups look different.

We now test if there is a statistically significant difference between the two groups of survival times. The default output, shown in Fig. 19.2f, is a non-parametric test that makes no assumptions about the underlying distribution for the survival times. See Collett (1994) page 40 for the derivation of this test.]
Fig. 19.2f  Test of difference between groups

Log Rank Test

Summary of results - logrank test

<table>
<thead>
<tr>
<th></th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>Expected</td>
<td>9.565</td>
<td>16.43</td>
</tr>
</tbody>
</table>

Logrank statistic 3.5
P-value (assuming chi-square under H0) 0.0608

As the P-value is close to 0.05, the logrank test gives some, though inconclusive, evidence that the survival curves of the two groups are different. The group with negative staining possibly has longer survival times, i.e. patients survive breast cancer for longer if they have negative staining. The graph, in Fig 19.2e, shows that the median is about 2 months, for the positively stained group. This compares to over 6 months for the other group.

The third stage is the modelling approach and this is particularly useful when extra data is available about individual patients, for example their age or whether they have had children. We may wish to eliminate the influence of these factors before testing the difference between the treatments.

For survival problems, this model-fitting stage needs the ideas of generalised linear models, and beyond Instat's capabilities. The statistics package SAS is the de-facto standard in the pharmaceutical industry and includes comprehensive facilities for modelling this type of data.

19.3 Time series analysis

We have already shown examples of exploratory graphs to examine time series, in Section 11.6, taken from Grubb and Robson (2000).

The first data set we use here is of the number of car drivers in the UK who were killed or seriously injured each month, from January 1970 until December 1984.

Fig. 19.3a Data  Fig. 19.3b Time for x-axis
The data are shown in Fig. 19.3a. Our initial task is to calculate a column to use as the x-axis for graphs. The Manage ⇒ Calc dialogue is shown in Fig. 19.3b, and the result is in column X5 in Fig. 19.3a.

Time series is a huge subject, on which there are many books, see for example Chatfield (1996). For illustration we here show a way of decomposing this time series into the “standard” components of trend, seasonality and residual. We then illustrate the same idea with other time series to show how individual parts of this analysis are often useful. Finally we show how the regression modelling ideas, described in Chapter 17, can be used to process this time series.

To decompose this monthly time series we first use a 13-point moving average to estimate the trend, in a way that allows the seasonality to be estimated as well. The values are

1/24 1/12 1/12… 1/12 1/24 or 0.04167 0.08333 0.08333 … 0.08333 0.04167

This is one of the standard options in the Statistics ⇒ Time Series ⇒ Moving Averages dialogue in Fig. 19.3c. The averages are saved into X6 and the residuals into X7. They are shown in Fig. 19.3f.

Fig. 19.3c Statistics ⇒ Time Series ⇒ Moving average

Fig. 19.3d Data

If X7 is plotted (not shown) it indicates the seasonal pattern is fairly constant over the 13 years. Hence a simple way of estimating this pattern is to find the mean for each month. This uses the Statistics ⇒ Data Summary ⇒ Column Statistics dialogue that was described in Chapter 12. It can be completed as shown in Fig. 19.3d. In addition to displaying the means, this dialogue also allows the fitted values, and “deviations” or residuals to be saved into further columns and they are needed here.

The results in the output Window are shown in Fig. 19.3e. They indicate that there are more accidents in the last quarter of the year, with December having the highest figures.
The fitted values, in X8, simply repeat the seasonal pattern for the whole record. They are shown, as the column called “Season” in Fig. 19.3f, with the residuals in X9.

The graph in Fig. 19.3g illustrates the decomposition of the time series into the different components. In terms of the data, shown in Fig. 19.3d it is a graph of x4, x6, x8 and x9 against x5. The key features of the trend appears to be a drop during the 1973-4 oil crisis and another drop, following the seat belt legislation that came into force in January 1983.

Finally in this part of the analysis we look to see whether there is any indication of serial correlation in the residuals. The Statistics ⇒ Time Series ⇒ Correlations dialogue is shown in Fig. 19.3h and the resulting graph is in Fig. 19.3i. It indicates that the first few autocorrelations may not be zero, though they are small. The reference lines in Fig. 19.3i are at (±2/√n), where n is the number of observations in the series. Very roughly,
autocorrelations outside these values are statistically significantly different from zero at the 5% level. We return to this point later.

Fig. 19.3h  Stats ⇒ Time ⇒ Correlations

Two further examples are shown briefly in Fig. 19.3j to 19.3m. Fig. 19.3j and 19.3k uses data from Dennett et al (1985). The data are a rainfall index for West Africa for each year from 1941 to 1984 and are in a worksheet called “waindex.wor”. Some of the data are shown in Fig. 19.3j to 19.3m with a smoothed line using a 7-point moving average, from Statistics ⇒ Time Series ⇒ Moving Averages. The graph of these data, shown in Fig. 19.3k, clearly indicates a downward trend in the rainfall index, since the late 1960’s.

The periodic component is the main feature of interest in the monthly maximum and minimum temperature data from Niamey in Niger. This uses the worksheet called “Niatemp.wor”, shown in Fig. 19.3l. The Statistics ⇒ Summary ⇒ Column Statistics dialogue, shown in Fig 19.3m for one of the columns, has again been used to give the monthly means and deviations.
Then there are two series to plot. One feature of interest, in Fig. 19.3n is the different pattern of the seasonal effect for the two series.

There are also some large autocorrelations (not shown) in the residuals, that are as high as 0.5. In Fig. 19.3o we show the dialogue to look at the cross-correlations. Here the zero-order correlation, shown in Fig. 19.3p is 0.66.
So far the fitting of these time series has been largely descriptive statistics. For example the weights used when calculating a moving average are arbitrary and up to the user to choose. More seriously without a functional form to the trend we cannot use the fitted models for forecasting.

Hence we now return to the accident data and use the regression modelling approach from Chapter 17 to fit a model to this time series. For illustration we first assume a linear trend and the same monthly seasonality as was considered earlier in this section. Our model can therefore be fitted using the Statistics ⇒ Regression ⇒ Simple with Groups dialogue, as shown in Fig. 19.3q. We first used this dialogue in Chapter 17, Section, 17.4

We show the results from fitting this model, in Fig. 19.3r, because they indicate how results could then be forecast. We would choose the corresponding month and extrapolate that line. For example, for selected months the results from the dialogue in Fig. 19.3q show the model is

Accidents = 1960 – 36.82 * (Year – 1970) (January)

Accidents = 1793 – 36.82 * (Year – 1970) (June)

Accidents = 2408 – 36.82 * (Year – 1970) (December)
Considering the trend more realistically could extend this modelling. One way of doing this would be to introduce another factor to distinguish between the time before and after the introduction of seat-belt legislation, i.e. January 1983. This could be done using a second factor in the model and we would then have to use the General Linear Models dialogue. Another possibility would be to use sine and cosine functions of the time of year for the periodicity, rather than simply a different level of the month factor.

Yet another possibility is shown in Fig. 19.3s. Knowing, from Fig. 19.3i, that the first order autocorrelation of the residuals might exist, we use the Manage ⇒ Transformations ⇒ One to One dialogue to produce a column with the data lagged by one month. We now use the General Linear Models dialogue, shown in Fig. 19.3s and add this new term in the model. This is called a first-order autoregressive model.

For January, the fitted model is as follows:

\[ \text{Accidents}(t) = 713 - 18.12 \times (\text{Year}-1970) + 0.5235 \times \text{Accidents}(t - 1) \]

### 19.4 Repeated measures

Our example is from Mead, Curnow and Hasted (1993), page 343. It is an experiment on bees. There were 9 hives and the brood area was measured on 5 occasions. The data are shown in Fig 19.4a and 19.4b.
We first consider the layout of the data, in preparation for the analysis. For simple studies the layout in Fig. 19.4a is appropriate. This is because the unit of interest is the hive. There are nine hives and we have taken five measurements on each hive. We usually take multiple measurements in any study and each measurement is a column. For example, if we had only measured at a single time point, but had recorded the area, number of bees, amount of honey, etc, then this layout would have been “obvious”.

But another way of looking at the data is to acknowledge there are two “levels” in the study. There is the “hive level” and there is the “hive at a time-point” level. The layout in Fig. 19.4b considers the data at the lower level. It is a simple case of repeated measurements, because there are five measurements on each of the hives, and they are at the same five time points.

Sometimes the data are not so simple and this “two-level” layout is needed. We described the general ideas in Chapter 9, Section 9.5. For example consider a study on babies growth in their first year of life. At the “baby level” we could have information like the age of the mother at birth, the parity of the baby, and so on. Then at the “baby at a time-point” level we have the age in months, the weight, etc. In that type of study we would be unlikely to have the same number of measurements on each baby and we would have to weigh them when they visit the clinic, so they might not be at the same age for each measurement.

The structure of the data in Fig. 19.4a and 19.4b typifies many studies. We can consider the data as constituting nine short time series, and it is therefore useful to look at the profiles with time. But the difference from Section 19.3 is that the main objectives are usually concerned with a comparison of treatment or other groups. The time series aspects are relatively incidental and are just the way the data had to be collected.

If you read Chapter 13 of this guide you might find the discussion of multiple levels to be familiar. There we looked at “multiple-response” data and had the same type of discussion of the different layouts of the data that could then be used.

There are many approaches to the analysis of repeated measures data. We mention just the simplest methods here.

There are two obvious starting points for the analysis. The first is a separate analysis at each time point. For this example that is a simple one-way ANOVA as described in Chapter 16.

The second is to look for each unit, here for each hive, at the data through time. This is sometimes called a profile plot. Here it is simple with a table or a graph, and both are shown.
below. The layout in **Fig. 19.4a** is fine for the table, for which the dialogue is shown in **Fig. 19.4c**.

**Fig. 19.4c** Statistics ⇒ Tables ⇒ Summary

**Fig. 19.4d** Graphics ⇒ Plot

In **Fig 19.4e**, we show the 5 time points vertically, because the pattern is simpler to see with this orientation. The table also shows the means for the three diets. We see that the pattern of the data is different for diet 1 compared with the other two diets.

**Fig. 19.4e** Tabular presentation of the bees data

<table>
<thead>
<tr>
<th>Summary</th>
<th>Diet</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>All</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>All</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Time1</td>
<td>368</td>
<td>308</td>
<td>284</td>
<td>320</td>
<td>624</td>
<td>288</td>
<td>488</td>
<td>490</td>
<td>444</td>
<td>260</td>
<td>390</td>
<td>368</td>
<td>389</td>
</tr>
<tr>
<td>Mean Time2</td>
<td>920</td>
<td>980</td>
<td>580</td>
<td>800</td>
<td>970</td>
<td>624</td>
<td>778</td>
<td>762</td>
<td>1330</td>
<td>290</td>
<td>828</td>
<td>815</td>
<td>803</td>
</tr>
<tr>
<td>Mean Time3</td>
<td>944</td>
<td>1014</td>
<td>754</td>
<td>904</td>
<td>803</td>
<td>570</td>
<td>636</td>
<td>672</td>
<td>1018</td>
<td>249</td>
<td>830</td>
<td>605</td>
<td>757</td>
</tr>
<tr>
<td>Mean Time4</td>
<td>1084</td>
<td>1008</td>
<td>904</td>
<td>1000</td>
<td>490</td>
<td>280</td>
<td>310</td>
<td>360</td>
<td>748</td>
<td>118</td>
<td>588</td>
<td>488</td>
<td>616</td>
</tr>
<tr>
<td>Mean Time5</td>
<td>1108</td>
<td>894</td>
<td>518</td>
<td>840</td>
<td>208</td>
<td>334</td>
<td>178</td>
<td>240</td>
<td>448</td>
<td>80</td>
<td>432</td>
<td>323</td>
<td>467</td>
</tr>
</tbody>
</table>

In Instat the layout in **Fig. 19.4b** is the route to showing the same information, i.e. the time profiles, in a graph, and the dialogue is given in **Fig. 19.4d**. The result is in **Fig 19.4f** and shows the slight problem with the graph, that nine series is quite difficult to sort out. In a real analysis it might be better to have three graphs, with one for each treatment. With a large study we might need a lot of graphs and the “trellis-plot” mentioned in Section 11.6 would be a neat way to look at the data.
The profiles, or previous similar studies, often indicate the next step. This is to calculate functions from the data at the time points that are useful in summarising the data across time. In this example Mead, Curnow and Hasted (1993) use two functions, namely

- Difference between areas at Time2 and Time1, i.e. \((x_4-x_3)\) from Fig. 19.4a.
- Slope of the simple regression line from Time2 to Time5.

We show these derived columns in Fig. 19.4g. Calculating the differences in X12 is simple, and uses the Manage ⇒ Calculate dialogue. However there is no special function to calculate the slopes of regression lines for each row of data. As explained in Chapter 8, statistics packages are essentially column calculators and usually have very limited facilities for row operations.

As a general problem there are two strategies to produce the results in such a case. The first is to find a way of doing the calculations in the package being used, i.e. in Instat. We solved the problem by writing a little macro. The way we did this is described in Appendix 1, when we look at Instat’s commands. Now that this macro exists it is simple to use. Just use Submit ⇒ Run Macro, as shown in Fig. 19.4h, and choose the macro called “rowreg” from
the Instat library. Then it will prompt you for the columns that will be used for the regression calculations, *Fig. 19.4i*, and where the results are to be stored, *Fig. 19.4j*.

*Fig. 19.4i  First prompt from macro*  
*Fig. 19.4j  Second prompt*

Once the slopes have been calculated, the method of analysis is the same as for the data at each individual time point, and here it is another one-way analysis. As a general concept, what we have done is to use our data at lower “level”, here the “hive at a time point” level, and calculated useful summaries at the hive level.

The second method is to derive the summary statistics prior to importing the dataset into Instat. We give an example in Excel and use a standard Excel statistical function to compute the slope of a regression line. Spreadsheet functions accept arguments horizontally, they are very flexible as to the layout of the data. In *Fig 19.4k* we have used the file called bees.xls. In this file we have added a row above the column names, storing integers that represent the day numbers. Then a new column is added, with the formula “=SLOPE(y’s,x’s)” as shown in *Fig. 19.4k*. Here there is a relative cell reference for y and an absolute one for x. This is then copied for each row.

*Fig. 19.4k  Results in Excel*

The results are in *Fig 19.4k* and are the same as from the Instat “rowreg” macro mentioned earlier. Now the data with these summary statistics can be imported into Instat for a one-way analysis on the nine hives.

We often find that these basic methods of analysis are sufficient for repeated measures data, because, unlike some of the more complex methods, they can easily be tailored to the objectives of the study. We describe the principles in a section called “Repeated measures made simple” in our guide titled “Modern Approaches to the Analysis of Experimental data”.

19.5 Circular data
There are a number of books on circular statistics, for example Batschelet (1981). We start by showing that the summary statistics dialogues, used in earlier chapters, are insufficient for processing circular data.

Consider a trivial sample of size two, shown in x1 of the data in Fig. 19.5a, where the data are recorded at 11pm and 2am. These observations might be the time of two calls to the police, or two occasions that a teenager returned home. What is the average of these two observations?

We do not need a computer to give the obvious, and correct answer as half past midnight, i.e. 00.30. But if we use a computer, then we would probably give the data on a 24-hour basis as shown in Fig. 19.5a. So our two values are 23 and 2, with a mean of 12.5, i.e. just after midday!

If we use Statistics ⇒ Summary ⇒ Describe, this is what we get, as shown in Fig. 19.5b. Instead we use Statistics ⇒ Summary ⇒ Circular, as shown in Fig. 19.5c. We set the option in the dialogue that the full circle is 24 hours. The results are in Fig. 19.5d.
When using the circular statistics in Instat we have to use simple units. So the results, in Fig. 19.5d are given in fractions of an hour, i.e. 0.5 is half past midnight, or 00:30. Instat does not have a special format for times.

With this simple example we could use the ordinary summary statistics, if we move the discontinuity in the data away from midnight. If the observation at 11pm is given as –1, i.e. one hour before midnight, then the ordinary mean of –1 and 2 gives the correct answer, of 0.5.

Measurements of variability are also different when considered on a circle. The standard deviation is similar in value and interpretation to that used in ordinary statistics. The circular variance is different, and is on a zero to one scale, with zero being no variation and one being the maximum possible. Here we see that the variance is 0.076, a low value, reflecting the closeness of the data points 23 and 2 on the circle.

For illustration we consider two further examples from Baschelet (1981). The first is of the time in the day that a city had traffic accidents and is shown in x2 to x4 of Fig 19.5a. The second is circular in space, rather than time and is in x5 of Fig 19.5a. It is of the direction of 15 homing pigeons.

The data are shown graphically in Fig. 19.5e for the accident data and Fig. 19.5f for the directions of the homing pigeons. A popular way of displaying circular data graphically is to use a rose diagram, which is like a circular histogram. Unfortunately this is one of the few facilities in the old Instat that is not yet available in this upgrade; so ordinary histograms have to suffice.
We see that the accident data are spread through the day, but with greater frequencies in the evening. The dialogue and results from the same Statistics ⇒ Data Summary ⇒ Circular are shown in Fig. 19.5g and 19.5h.

**Fig. 19.5g Statistics ⇒ Summary ⇒ Circular**

**Fig. 19.5h Results for accident times**

![Image of Descriptive statistics for circular data and Commands and Output]

We see, from Fig. 19.5h that the mean time for accidents is about 4:30pm and the standard deviation is over 5 hours. That large standard deviation corresponds to the histogram, shown in Fig. 19.5e, which showed the accidents spread throughout the 24 hours.

The dialogue and results for the homing pigeons are shown in Fig. 19.5i and 19.5j. We see that the average direction is 303 degrees, with a comparatively smaller spread of 26 degrees.

**Fig. 19.5i Statistics ⇒ Summary ⇒ Circular**

**Fig. 19.5j Results for directions**

![Image of Descriptive statistics for circular data and Commands and Output]

In the last dialogues we have also included the option to fit a simple model. The von Mises distribution is the equivalent of the normal model in the analysis of circular data. It is symmetrical and also has two parameters.

We do not continue with the analysis, but note that, with the fitting of this distribution, we have moved from purely descriptive methods, covered earlier in Chapters 11 to 13 of this guide, into the modelling ideas, introduced in Chapters 14 and 15.

In the Climatic Guide, Chapter 12, Section 3, we consider a further example, which is of the time of day that it starts to rain. There we take a further step, and ask whether there is any
relationship between the time the rain starts and how much rain falls. This uses the regression ideas from Chapter 17 of this guide.

19.6 In conclusion

We have stressed the value of simple approaches in this chapter. Often, however we need to move to models, if we are to analyse the data fully. This is the parallel of moving from Chapters 11 to 13, into the regression ideas of Chapter 17.

For survival studies, described in Section 19.2, the different types of model are described by Collett (1994), together with information about the facilities for fitting these models in the standard statistics packages.

For the time series, described in Section 19.3, Chatfield (1996) discusses the use of different modelling approaches. He suggests that a relatively empirical model, called Holt-Winters, is often useful for those who want a routine method for processing time series, particularly if they are interested in forecasting. This is simple enough to be feasible in a spreadsheet, such as Excel and is described in Berk and Carey (2000). The comprehensive family of models, called Box-Jenkins models, is in the powerful statistics packages, and also needs more user intervention to choose the appropriate models.

There are different approaches to modelling the repeated measures data we described in Section 19.4. Some packages simply include a “repeated” option and this usually fits quite a general model. In contrast, Fig. 19.6a shows the options for repeated measures in Genstat, and this suggests a range of different approaches. For example, one option is called “antedependence analysis”. Where models are to be used, this approach is sometimes attractive, because it explicitly models the repeated measures as simple time series.

Fig. 19.6a  Approaches to repeated measures in Genstat

In each of Sections 19.2 to 19.4 there are complications if models are to be fitted. In Section 19.2, on survival, we need methods to cope with the analysis of non-normal data. In time series, introduced in Section 19.3, one key point is that the successive observations are correlated with each other. And in the repeated measures discussion we saw that we have data at multiple levels.

The powerful packages have general facilities for analysing multi-level models, and these can cope with inter-correlations between observations. These are usually called mixed
models. For example, this is PROC Mixed in SAS, or REML in Genstat. These are now being extended to cope with non-normal data.

Some of the teaching ideas are well illustrated by the analysis of circular data, described in Section 19.5. This subject will be new, even to many statisticians. However, not much is new, even in a subject that initially seems so different. We still need descriptive methods, even if the summary statistics have different formulae. And simple modelling, followed by correlation and regression ideas to study relationships is as before. What is new is the capability to give a practical overview of the subject, to non-statisticians.

Unlike the other topics considered in this Chapter, circular and directional statistics is not covered in a standard way by the major statistics packages. Macros have been written for some, like Stata and S-Plus. And there is also a specialist package, called Oriana, for processing such data.
Chapter 20 - Collecting random samples

20.1 Introduction

Pseudo-random numbers have many interesting applications. Much can be learned about the behaviour of a probabilistic model by getting the computer to simulate data according to the 'rules' of the model. Sometimes a theoretical analysis is too difficult to contemplate and simulation may then be the only way to study the model.

Another application arises in data analysis. We have earlier described probability plotting as a means of checking whether data are consistent with some theoretical distribution, and usually we produce a plot which should 'roughly' be a straight line if the distribution fits the data. But how 'rough' can the line get before we start to have doubts? A good procedure is to generate several samples of pseudo-random data, with the same sample size as your data, and subject these samples to the same probability plot. In this way you can get a feel for just how much deviation from the 'ideal' straight line you can reasonably expect if the model were a good one.

Computer-generated random samples are also useful in teaching. Although no substitute for learning by analysing 'real' data, there is no better way of understanding the nature of sampling variation than by getting the computer to repeatedly generate samples for you.

20.2 Generating random samples for teaching

We first use File ⇒ New to give a worksheet with 100 columns of length 150 and then use the Statistics ⇒ Random Samples dialogue, as shown in Fig 20.2a.

In the dialogue in Fig 20.2a we have specified we would like samples of size 9. The number of columns to be generated is dictated by the number we give to store the data. Here we have asked for 40 columns. On the right-hand side we have asked for samples from a normal distribution with mean 20 and standard deviation 6. The data for the last 3 columns are shown in Fig 20.2b.

In the dialogue we have also asked for summary graphs. The effect of this is to produce the columns x41 and x42, also shown in Fig 20.2b. They contain the 40 means and medians, which are plotted as shown in Fig 20.2c and 20.2d.
These histograms are designed to support the teaching of simple inference, as discussed in Chapter 15. The graph titles also show the theoretical values. So in the first graph, the sampling distribution of the means should have a mean of $\mu = 20$ and a standard deviation of $\sigma/\sqrt{n} = 6/\sqrt{9} = 2$ in this case. We see the actual values are close. We also see that the sampling distribution of the median has a greater spread, which is why the mean is used as the conventional estimator.

In Fig. 20.2a a further option is to plot confidence limits and Fig. 20.2e shows this plot for another set of 40 samples. This graph is designed to support the teaching of confidence intervals, another topic mentioned in Chapter 15 and in the good-practice guide concerned with statistical inference.
In Fig. 20.2a, the "seed" field can be used if you had wanted to generate the same 40 samples again. This would then give the equivalent graph in Fig. 20.2e to the histogram in Fig. 20.2c. Give any integer you like as the seed, and then give the same value again the next time.

In Fig. 20.2a we chose to sample from the normal distribution, but other distributions can also be used. For example, if you use the exponential distribution, then the same graph as shown in Fig. 20.2c illustrates the central limit theorem.

20.3 Simulation as a research tool

Statistics packages provide ideal tools for simple simulation studies where the task involves generating columns of data. When individual values only are needed, then the flexibility of a spreadsheet may make it a more convenient environment. When the simulation model is complex a special program may be written, either in a standard language, such as Visual Basic, or in a simulation-modeling environment, such as Stella. And recently, the speed of modern computers has enabled some statistical methods to advance dramatically, through simulation. Perhaps the most exciting is the whole area of Bayesian statistics, see for example Gelman et al (1998), Congdon (2001) and also the Winbugs software.

Instat can be used to illustrate simulation ideas, particularly for teaching purposes, as we show below. Some simple examples are in the Case Study called "Using Simulation to Avoid Messy Algebra". As a further simple example we consider the simulation of the occurrence of rain for 50 years, where the chance of rain each day is 0.4.

We start with a new worksheet that has 120 columns each of length 100. We then use the Statistics ⇒ Random Samples dialogue, as shown in Fig. 20.3a, to give the simulated data, part of which is shown in Fig. 20.3b. In completing the dialogue we have chosen to set the "seed", i.e. the starting point for the random number generator, to 3456. This can be replaced by any value. It is not needed in this first step, but we will use it later.
We suppose that our initial objective is to estimate the chance of long dry spells, i.e. no rain. In particular we need to find the chance of a long dry spell within a 30-day period. We will consider “long” to be at least 5 days and at least 7 days.

We can find the length of the longest spell in each column using Instat’s **Climatic ⇒ Events ⇒ Spells**, as shown in **Fig. 20.3c**. We start from day 11, so the “system” has enough time for the spell to start before the beginning of our “month” of interest. The dialogue then gives the length of the longest run of zeros in the data from row 11 to row 40.

We have chosen to put the summary data in x101. Some of the results, i.e. the longest dry spell in the 30 days for each of the 50 years, are shown in **Fig. 20.3d**. By inspection of x101, or by using the **Statistics ⇒ Summary ⇒ Column Statistics** dialogue, we find that 22, i.e. 44% of the years have a dry spell of at least 5 days, while just 28% of the years have a dry spell of 7 days or more. If you have used a different starting seed, or no seed at all, then you will get different answers, but they should be close to these results. (They would be very close if we had been able to simulate with 5000 years, rather than 50.)
Now we suppose that our main task is to investigate whether the risk of long dry spells changes much, if there is climatic change.

To start this investigation, we repeat the exercise, but with a reduced probability of rain of 0.35. We use the same starting seed, for reasons that we explain below. The data for 50 years are put into x1 to x100 as shown in Fig. 20.3e. Then we produce the 50 longest spell lengths for this situation in x102. In Fig. 20.3f we show some of these data, together with the column of differences between x102 and x101.

In Fig. 20.3f we have also added a column, x104, with the numbers 1 to 50. This is used in Fig. 20.3g to give a graph of the two results, as shown in Fig. 20.3h.

The graph shows the advantage of using the same random number seed in this simulation. The same sequence of random numbers has been used in each case, and hence the years are “paired”. This makes it easier to look at the differences between the two situations. We have kept the individual pairs of years, x1 and x51, etc. to illustrate this point further, though this would not usually be needed. If you inspect them you should see that x51 might be dry (i.e. have the value 0) when x1 has rain, but the converse never happens. So, the graph shown in Fig. 20.3f and the data in x103 are always zero or positive. This makes it possible to ask conditional questions. For example, if we select the 36 years where there was no dry...
spell of more than 7 days in x101, we can find that there was a problem in 8, i.e. in just over 20% of those years, when the chance of rain was reduced.

The use of the random number seed is an illustration that when we simulate we have total control and should use that control as effectively as we can. Here there is just one “parameter”, namely the chance of rain, that is varied, but usually there will be many. To study the pattern of responses we can consider the simulation as an “experiment”, and each parameter of interest is a “factor” that we wish to study.

Hence the design of a simulation study is often like the design of a factorial experiment. The analysis will therefore often use the same ideas, shown in Chapter 16, for analyzing experimental data. Where other software is used for the simulation, it is still possible to transfer the summary results to a statistics package for the analysis, just as one would do for ordinary experimental data.

Most simulation studies involve, as here, a well-defined sequence of steps. If a statistics package is being used, then (apart from teaching purposes), it will usually be inefficient to use dialogues for these
Chapter 21 - Designing the Study

21.1 Introduction

We return to the starting point of design, having seen, in earlier chapters, how the data can be analysed. Many of our "good-practice" guides are concerned with the planning stage. We list them in Fig. 21.1a.

<table>
<thead>
<tr>
<th>Area</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview guides</td>
<td>Statistical Guidelines for Natural Resources Projects</td>
</tr>
<tr>
<td></td>
<td>Projects Combining Quantitative and Qualitative Survey Methods</td>
</tr>
<tr>
<td>Planning guides</td>
<td>Some Basic Ideas of Sampling</td>
</tr>
<tr>
<td></td>
<td>Guidelines for Planning Effective Surveys</td>
</tr>
<tr>
<td></td>
<td>Sampling and Qualitative Research</td>
</tr>
<tr>
<td></td>
<td>Concepts Underlying the Design of Experiments</td>
</tr>
<tr>
<td></td>
<td>On-Farm Trials - Some Biometric Guidelines</td>
</tr>
<tr>
<td></td>
<td>One Animal per Farm</td>
</tr>
</tbody>
</table>

Important though they are, we omit any detailed discussion of the initial planning steps. They include the formulation of the objectives and decisions on the type of study to be undertaken that can meet those objectives.

We often find that these initial steps have been decided, before a statistician is involved. For example a cross-sectional survey is to be undertaken. We then meet the client and are merely asked “How large a sample do I need?”

Equivalently, in an on-farm trial we may be asked “How many farmers are needed?” and “Do we need to have replicates within a farm?” We address the issues of sample size in Section 21.2.

When statisticians have been involved as full team members, i.e. right at the start, they can contribute to the discussions on the most effective types of study. We sometimes find that researchers only consider types of data collection for which they have been trained, and it can then be difficult for them to meet their stated objectives.

As an example we consider a study on cotton in West Africa. The problem was to help farmers, who felt that the rainfall pattern was changing, and therefore wanted to know the new recommended planting dates for cotton. They asked an experimenter, who designed an on-station trial that investigated alternative planting dates. It ran for three years.

One alternative would have been to examine the monitoring data, in this case rainfall records, that could have been made available. A second possibility would have been to plan a survey, or a participatory study, to gather information from farmers. Perhaps a combination that involved all elements might have been more effective. The researcher certainly found that the on-station trial, by itself, did not satisfy the farmers’ objectives.

In Chapters 5 and 10 we have already given our views on the incorporation of design issues in the teaching, both for statisticians and in service courses. In Reading this is partly through the statistical games, described in Chapter 5, that permit students to practice the process, from planning the design to reporting the results within limited time periods.

Once we know the design, then statistical software can help by providing a randomisation, as we show in Section 21.3. We then describe, in Section 21.4, how some aspects of the proposed design can be evaluated before the data are collected.
21.2 How large a sample do I need?

Sample Size

Statisticians are often asked to advise on the size of the sample needed for a given study. Here we first show the results from using the formal calculations and then we indicate how these results can be used in practice.

An example is taken from Lemeshow et al (1990). A study is planned to test whether a dietary supplement for pregnant women will increase the birth weight of babies. From a pilot study the standard deviation of birth weight is estimated to be about 500g. The scientists would like to have a good chance, say 80%, of detecting a mean difference, if it exceeds 100g. This chance is known as the “power” of the test.

Use the Statistics ⇒ Sample size dialogue, and complete it as shown in Fig 21.2a.

The results are shown in Fig 21.2b and indicate that a sample of 393 subjects is needed in each group.

There are many reasons why this calculation should be viewed as just a part of the discussion on the design of the study. In this case let us suppose the scientist finds that this sample size exceeds her budget. The largest sample that her resources will allow is about 150 subjects per group. The same dialogue, as above can then be used again, but we now specify the sample size and calculate the increase. The results in Fig 21.2d show she has a good chance of detecting the benefits of the supplement if the mean increase in birth weight exceeds 229g.
The final way of using the same dialogue is to ask what is the change that a true difference of just 100 g will be correctly detected, i.e. declared statistically significant at the 5% level, if 150 subjects are used. As shown in Fig 21.2f, the result is 0.231, so the chance is just 23% and not 80%, as it would be if the resources had permitted recruiting 393 subjects.

Often the sample size of a study is determined more by the financial resources and time constraints than by the result of the calculations using this dialogue. This is the parallel of the fact that the analysis also can rarely be reduced to a single calculation. Most studies have multiple objectives and make more than one measurement.

This complexity does not negate the value of the calculations, but it indicates that the calculations will be just a part of the planning process. It is an important part, because it is pointless to undertake a study if the major objectives are unlikely to be met. Similarly, it is
waste of resources to collect a huge sample if the objectives can be met with a much smaller
sample, i.e. fewer subjects in this context.

Instat only includes the calculation of sample sizes for the differences between two means or
between two proportions. Sometimes these facilities can provide guidance, even when the
study is more complicated. For example, in Chapter 16, Section 6, we considered an
experiment with 16 treatments, with the treatment structure being a 2*2*2*2 factorial. There
were two replicates overall and a key result was that the effect of applying nitrogen was
important. Suppose that a similar trial is planned elsewhere and we would like to detect the
effect of nitrogen, if it increases mean yield by more than 7.5 kg.

We will assume roughly the same variability as in the previous trial. There, Fig 16.6j, the
residual mean square was 24.268, so the standard deviation is about 5. Use of the same
formula with power at 80% shows that a difference of 7.5 units needs 7 replicates.

Now, in the previous trial, there were no significant interactions with nitrogen. So with two
overall replicates, plus the “hidden replication” that comes from averaging over the other
factors, we have a total of 16 replicates and this is ample. It may be that, this time, there will
be some 2-factor interactions, and then there would only be 8 replicates, but that is still
sufficient. Hence this design seems sensible, at least for this objective.

Most of the standard statistics packages, including Minitab, offer a wider range of scenarios
for the calculation of sample size. There are also some specialist packages, for example
nQuery and Pass.

### 21.3 Choosing random samples

We suppose there are 20 people in a class and a random sample of 6 are to be chosen. To
show how this is done, start with a new worksheet and put the numbers 1 to 20 in the first
column. The Manage ⇒ Data ⇒ Regular Sequence can be used for this. In preparation for
the second task also use this same dialogue to put 1/20, i.e. 0.05 into the next column. The
worksheet should therefore be as shown in Fig. 21.3b.
Now use the Statistics ⇒ Random Samples dialogue, that we used already in Chapter 20. This time the “Sample type” is specified to be a “Data column (without replacement)”

We have generated 3 random samples. Our samples are displayed in the output window, and also in columns x3-x5, as shown in Fig. 21.3d. Notice that the six numbers are different within each column, because we have sampled without replacement. They may, of course, repeat in the different samples.

If you needed 3 samples of size 6, and did not want any repeats at all, then you would have to generate a single sample of size 18, and take successive groups of 6 from this column.

Sometimes you do want the units in the sample to be able to repeat. Then we use the second column, shown in Fig. 21.3b and 21.3d. In the same dialogue, shown in Fig. 21.3e, we use the option to sample from the data column, with replacement, and give x2 as the column that provides the relative probabilities. In this case they are all equal, but they do not have to be. For example, we may wish to sample villages in proportion to their size, or children in proportion to the number of raffle tickets that they bought. The column, x2, in our case, does not need to add to 1. It just provides the relative probabilities of selection.
The random samples that we generated are in x6 to x8, in Fig. 21.3f. We see that unit 11 is repeated in x6.

An alternative way of sampling with replacement, if the probabilities of selection are equal, is to sample from the uniform distribution. We give an example in Fig. 21.3g for the same problem. The results are shown in the output window in Fig. 21.3h and displayed without decimals, in Fig. 21.3f. We used the uniform distribution from 0.5 to 20.5, so that the rounded values would parallel the previous method.

If you just want a random permutation of the integers 1 to 20, this is the fourth way of sampling provided in the Random Samples dialogue. The dialogue is shown in Fig. 21.3i, and part of the output window is in Fig. 21.3j. This would also be a way of sampling without replacement, by just choosing the first 6 values in each permutation.
21.4 Will the sample be effective?

Once we know the sample size and have a sampling scheme it is often useful to generate some “pretend” results. Then we can look at the structure of the tables we will produce later and check that they seem satisfactory. For example, suppose a survey is planned on smoking by different groups of people. We assume that a sample of 30 people is suggested and a key table is to be of “age by smoking”.

The last alternative that we illustrate in this section is to sample from factor columns. As an example we consider a trial with 8 treatments that we want to randomise and need 3 replicates. We enter the column in non-randomised order and define it as a factor, attaching the appropriate labels. Then we use the option to sample without replacement, as shown in Fig. 21.3k. The original order and the resulting randomised columns are shown in Fig. 21.3l. An alternative way is to permute the numbers 1 to 8. Then to present the results as shown in Fig. 21.3l, make each of the columns into factors and attach the labels.
There are to be 4 age categories and 5 smoking, as we have entered into $x_1$ and $x_3$ of Fig. 21.4a. We then enter the appropriate label columns as shown in Fig 21.4b, and then use the Manage $\Rightarrow$ Data $\Rightarrow$ Factor dialogue to make $x_1$ and $x_3$ into factor columns, as shown in Fig 21.4d. We can now ask the person responsible, what are the likely proportions in each category. In the absence of further information we have assumed each category to be equally likely and have therefore entered columns of 1’s into $x_2$ and $x_4$ in Fig 21.4a.

Now we use the Statistics $\Rightarrow$ Random Samples dialogue, as in Section 21.3, to generate random data. The dialogue is shown in Fig. 21.4c and has been used to generate the data in $x_5$ and $x_6$ of Fig. 21.4d.
Now use the **Statistics ⇒ Tables ⇒ Frequency** dialogue, as shown in **Fig. 21.4e**, to give a table of counts like **Fig. 21.4f**. The actual contents will differ for each random sample, but the general structure will be the same. The actual numbers are unimportant, because they are not real data, but the structure of the table can form the basis of a discussion on the required sample size. In this example, if the 2-way table is important, then we see that a larger sample than 30 is needed.

It is sometimes possible to check on potential problems in a sampling scheme or an experimental design after some, but not all the data are collected. As an example we consider the rice survey that was described in Chapter 3 and which has been used in many chapters. We suppose that most of the data have been collected, but we have yet to record the yields. This is common in practice. Thus we interview the farmers at the start of the season and record the variety they use, and fertiliser applied, etc. Then, at the harvest time we measure the yields.

Hence the **Statistics ⇒ Tables ⇒ Frequency** dialogue, shown in **Fig. 21.4g**, could be used after the first interview. We notice, from **Fig. 21.4h**, that the data are very unbalanced. Had this been a trial, rather than a survey, we might have tried to have 3 farmers with each of the 12 “village by variety” combinations. For us the replication ranges between zero and seven.
If a key aim is to evaluate the difference in yields for the three varieties, taking account of the different villages, then a small extra study, particularly of farmers using the “New” variety might help this to be realised.

The other potential problem with the sample shown in Fig. 21.4h is the lack of balance. The extent to which this is a problem can be investigated with the Statistics ⇒ Analysis of Variance ⇒ General dialogue, as shown in Fig. 21.4j, to estimate the “contrasts” shown in Fig. 21.4i. This dialogue has a special option for a “dummy” analysis.

The results are shown in Fig. 21.4k. We see that our two contrasts are at least 90% efficient. Hence the lack of balance does not cause a serious problem.

In Instat the other statistics dialogues do not have a special facility for a dummy analysis. But it is always possible to generate a random sample to investigate the form of the analysis and use this to consider the effectiveness of the design.

21.5 In conclusion
Most statistical packages offer some support on the design of experiments and surveys. A dialogue on sample size for simple problems is common and many offer extensive facilities for designing sampling schemes for surveys and experiments. They differ in emphasis. For example Minitab has facilities that are particularly targeted at industrial experiments, while Genstat has dialogues for the designs that are common in agricultural field trials.

The facilities for dummy analysis are also possible in some packages, while all allow the generation of random samples.
Chapter 22 - Finally

22.1 Looking back and looking on

Part of the impetus for this new version of the Introductory Guide has been an evaluation of the role of statistics in agricultural research in a number of Universities outside the UK. MSc students in 9 universities were taught statistics within their course and then undertook a substantial research project, jointly with members of staff. At subsequent presentations of the research it was clear that the training in statistics had not provided the students with all the statistical “tools” they needed to support their research project.

In meetings with different stakeholders to try to understand the problem, the agriculture students suggested that changes are needed to the courses in statistics at undergraduate as well as postgraduate level. Staff also asked for training in statistics.

The providers of the training were sometimes statisticians, and sometimes they were agricultural researchers who were interested in statistics. They were aware of the problems and, in some Universities they had already begun a process of substantial change, e.g. Akundabweni (2000).

Our work at Reading with groups, both inside and outside the university, indicated that the problems above were common everywhere. The lack of confidence, even fear that many non-statisticians still have about statistics provides a message to trainers that we have not been very successful in our teaching. Students may pass exams, but they often still do not understand the key concepts, nor are they particularly adept at statistical aspects of their subsequent project work.

This guide is therefore an attempt to contribute to a solution. We have used these ideas in our own training, sometimes with Instat, and also using other software. The approach is not original. It is consistent with the message that is given in many popular statistics textbooks such as those listed in the references to this guide. What is perhaps new is the ease with which the ideas in this guide can be put into practice, within the time and framework for current courses in statistics. This follows from the computing experience of most students and the ease of use of modern statistical software. Hence we can incorporate the computer fully into our teaching of statistics, without diverting the training into a course that teaches (statistical) computing, rather than statistical skills.

22.2 More statistics

This is just an introductory guide, and Instat is just an introductory package. Hence we have used the last section of most chapters to outline further work that is possible. Sometimes this further work is possible with Instat, but more powerful software is often needed.

The main messages on data analysis that we have emphasised here include:

- Be prepared to spend time organising the data. This will pay dividends in enabling a variety of analyses to be done effectively.
- Look at the data critically at all stages in the analysis.
- In many studies most of the analyses can be descriptive. Just because training courses spend most of the time on statistical inference, does not mean that most analyses need do the same.
  In this guide Chapters 11 to 13 were purely on descriptive methods, and later chapters included a mixture of description and modelling
- Recognise the role of statistical inference, or modelling. The concepts were introduced in Chapters 14 and 15, and were then applied in later chapters.
Recognise also the limited role of significance testing. While the ideas are useful, we often find significance testing is over-used and sometimes has distracted researchers from alternative analyses that correspond more to the objectives of their study.

For those who would like more we emphasise three topics that we consider to be of general importance.

The first is generalised linear models. This has allowed all the ideas of ANOVA (Chapter 16) and regression (Chapter 17) to be applied more generally, i.e. for a wide range of models when the data are not from a normal distribution. One example was described in Chapter 18, when we considered log-linear models.

The second is multilevel models, i.e. the analysis of data at more than one level. The methods developed to handle such data can also be applied in many other areas such as spatial models. These methods are available now in the more powerful statistical packages, and again are both for data from normal models and more generally. If your favourite statistical software does not include these features, or anyway, look also at the package called ML-Win that is designed to fit multilevel models.

Our third area is Bayesian methods. One topic is called Bayesian belief networks and this is already used extensively. See the Netica software if you are curious. The general area of Bayesian inference is not in the standard statistics packages. The book by Gelman et al (1998) shows the extent of recent progress in Bayesian data analysis and the availability of software, including Winbugs, is helping to turn this subject from one of academic, to one of great practical interest.

### 22.3 Making more use of the software

We believe that we now have a window of opportunity to improve statistics teaching and hence the analysis of data. This has been provided by three recent developments:

- The statistical packages are now very easy to use.
- It is also easy to transfer data and results between packages.
- Users are comfortable with the computer.

In the past we have spent a considerable proportion of the time in courses and text books on the theory. Courses have also spent a long time on the mechanics of data analysis with a particular package. The software and the computing experience of users has provided the opportunity for training to spend more time on a broader type of course. It is this broader training that is implied by the materials and approaches described in this guide and in the texts that are in the list of references.

There is a price we may be paying for the ease of use of modern software. A new generation of users is avoiding learning any of the “joys” of programming. Programming, or at least typing commands was the way that computers were used until recently. We see part of this price when watching an Excel user who spends endless hours using Copy/Paste, because he knows nothing about the use of Excel macros. Some user perhaps do not even realise that computers were built to do the repetitive tasks. Now they are doing all the repetition and their powerful computers are sitting almost idle.

The same applies to the use of statistical software, and perhaps particularly to the powerful statistics packages. We propose that users need to be able to type commands, if they are to use the software to its potential. Perhaps we still need to teach just a little of the concepts of programming to help such users, to overcome their fear of typing commands. This applies at least to those users who are offering their colleagues some support.

What is different is the point at which we would teach these programming ideas. Earlier they were taught at the start, when the software was introduced. This meant that learning the
Software was mixed with its use to support the statistical work. Now learning the language can be a small task at some point, when the software is already familiar. We illustrate in Appendix 1 with Instat, but the same applies to all statistics packages.

### 22.4 Putting the jigsaw together

It is now many years since the important paper of Nelder and Wedderburn (1972) introduced the subject of generalised linear models. That work was important, partly because of the way that our ideas of regression modelling, described here in Chapter 17, could be applied to a wide variety of other types of data, such as the log-linear models, described in Chapter 18. But also important was the availability, at the same time, of the general computer software, called GLIM (GLIM stands for Generalised Linear Interactive Models) that was developed at Rothamsted. That package allowed statisticians to try these new models and extend still further the concepts that were encouraged by the new approach.

In GLIM, the "I" in the name of that package was also innovative, because it encouraged us to work interactively, trying different models, much as is natural now that computers have become so easy to use in an interactive way. At that time, it would still be 10 years before microcomputers appeared, and we usually used computers by providing them with a deck of cards, containing both the data and the instructions for the analysis. However, early though it was in computer days, some of the main statistics packages, that are currently still the market leaders were already in common use.

Now these statistical and computational ideas, and others, are available for everyone to try. It is up to teachers of statistics to make the subject accessible. It is then for students and practitioners to overcome their fears of the subject, and to realise the benefits to their work of trying to exploit their data fully. We often find that much time has been spent on the gathering of data, compared to the time that is devoted to the analysis. We hope that some of the ideas in this guide will help users to understand how to look constructively at their data and, perhaps even to enjoy the time they spend on analysis a little more.

The problems we have tried to address in this guide relate to the fear and poor understanding that many people still have of statistics. The problems are well known, but they persist, despite efforts from statisticians to adapt their teaching and from recipients to understand the subject.

We must not underestimate the complexity of the problems in teaching and learning statistics. If there were a simple solution, then it would have been found and adopted a long time ago. From the persistence of the problems we deduce that making improvements will be a complicated process. For example, many of the ideas in this guide were used to redesign a training course that we currently give to postgraduate students at Reading. There are two main groups, one of which seems to appreciate the changes and the relevance of the new course. The other does not!

Hence, while we feel we have a window of opportunity, we must be realistic. We hope that this guide can be at least a piece of the jigsaw, to help both trainers and trainees in statistics.
Appendix 1 - Using commands as well as dialogues

A1.1 Introduction

In Chapter 4, Section 4.5, we explained how you use Instat by typing commands, instead of just using the dialogues and also why you might want to do this. In this appendix we look further at Instat’s language.

A1.2 Instat command lines

In this Section we describe the basic “rules” or “grammar” of Instat’s language. To explain how Instat processes commands we assume that you have a worksheet with some data in x1, x2 and x3 (x4 is explained below), such as is shown in Fig. A1.2a.

Fig. A1.2a Simple worksheet

We assume here that you are typing in the “Commands and Output” window, as shown in Fig. A1.2b. The command, shown in bold on the first line, could alternatively be generated using the Statistics ⇒ Summary ⇒ Describe dialogue. The Manage ⇒ Calculations and Manage ⇒ Remove, dialogues could be used for the others. The three commands shown in Fig. A1.2b, are as follows:

\[
\begin{align*}
: & \text{describe x1} \\
: & \text{calc x4} = x1 + x2 + x3 \\
: & \text{remove x4}
\end{align*}
\]

Fig. A1.2a was captured at this point.

When you type into the “Commands and Output” window, Instat is an interactive language. You type a line of text. When you press <Enter> this line is submitted to the Instat server.

The program reads the line, looks for a command in the text, searches its dictionaries for the command, checks for errors and, if there are none, executes the command. The results are sent back to the “Commands and Output” window.

Instat then waits for the next command. If an error is found, then the command is abandoned and an appropriate error message is displayed.

This is why the window is called “Commands and Output”. You can type commands into the window, and you also see the results there.
The prompt given by “:” indicates that Instat is ready to accept a command. Instat only uses the first 3 characters of the command name. Apart from any arguments (columns, numbers, filenames etc.) that the command may need, all other text in a command line is optional and purely for the user's benefit.

The Instat command must be the first word in a command line. Commands and text may be in any mixture of upper and lower case letters. For example, the following four commands are equivalent.

: DESCRIBE THE DATA IN X1
: Describe X1
: desclobinatex1 youstupidcomputer
: des x1

Besides the command prompt “:”, other prompts given by the program include “sub:” when a subcommand is expected and “data:” when data are input from the keyboard.

More than one command may be entered on the same line. A colon is then used as a separator. For example, : INF : DIS X1-X4

Columns are denoted by Xn, where the letter X (or x) must be typed directly in front of the column number (n), e.g. X12. There must be no spaces or other characters between X (or x) and the column number.

Constants are denoted by K1, K2, ...; Label Columns by L1, L2, ..., and Strings by S1, S2, ... See Chapter 6, Section 6.5, for a description of the data types available in Instat.

Some arguments of a command need a special symbol in front. This is to distinguish them from the arbitrary text that Instat ignores. Thus, when a filename, or a directory is included in an Instat command, it must be prefixed by @ with no spaces between @ and the filename; for example @RAIN, @SURVEY.

If you need some text as part of a command, then it is put into quotes, for example

: plot x1 x2;title "A simple scatterplot"

You may refer to columns by their name. These are preceded by a single quote, and the name may be abbreviated, as long as it is unique. When looking for names, Instat is not case sensitive, so 'Yield' is the same as 'YIELD' or 'yield'. For example

: name x2 'Yield'
: des 'yield

The closing quote is optional.

If you give a new name when the column is being written, then Instat will automatically use the next free column, and give it that name. For example, if x5 is the first free column, then

: calc 'tot = x1+x2
: calc x5 = x1+x2
: x5=x1+x2

are all the same to Instat. The last version also shows that the calc command is special, in that the command name is optional. Another special command in this way is when you want to use Instat as a calculator, but not store the results. So

: show 3+4 : show x1+x2, are the same as
: ? 3+4 : ?x1+x2
Many of Instat's commands have subcommands which modify the behaviour of the command. The idea is that it is normally possible to use a command in a simple way, with default options chosen by the program, but should you need more flexibility, then subcommands are available. Subcommands are separated by a semicolon. Here is an example:

: des x1;median ;percents 20 50 80

The command name is unique. For example there is just one Instat command that begins with “des” and that is “describe”. The subcommands for any command are dictated by which command you use. The Reference Guide gives the names of all Instat commands, and the allowed subcommands for each of them.

If a group of columns with consecutive column numbers are used in a command, then you can abbreviate the list. For instance,

: des x1-x3  is equivalent to
: des x1 x2 x3

The same abbreviation of the syntax can be applied to the other types of data.

If, because of a mistake in typing in a command, it is not executable, Instat will give an error message. For example,

: GEScribe X1-X3  (Problem: describe spelt wrongly!)

No such command  (Instat's response)

: Display  (Problem: Does not say what to display)

Command too short

: DEScribe X3  (Problem: Suppose we forgot to enter the data first!)

No data in X3

No doubt you will discover other error messages while you are typing Instat commands. We hope that their meaning is clear, although they are rather terse!

A1.3 The Instat commands

There are about 100 words in the Instat language. They are shown below, divided into topics.
## Fig. A1.3a Instat’s commands

<table>
<thead>
<tr>
<th>Data entry editing and display</th>
<th>Managing data</th>
<th>Files and macros</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCuracy</td>
<td>CALcule</td>
<td>CLOse worksheet</td>
<td>BOXplot</td>
</tr>
<tr>
<td>CALcule</td>
<td>COPy</td>
<td>CREate worksheet</td>
<td>DEFault</td>
</tr>
<tr>
<td>DELete</td>
<td>ENTer</td>
<td>EXECute macro</td>
<td>HIStogram</td>
</tr>
<tr>
<td>DISplay</td>
<td>FACTor</td>
<td>EXIt (from macro)</td>
<td>LINe</td>
</tr>
<tr>
<td>ENTer</td>
<td>GROup</td>
<td>INFo</td>
<td>PLOT</td>
</tr>
<tr>
<td>FIXed</td>
<td>INDicator</td>
<td>LOOp</td>
<td>REPLOT</td>
</tr>
<tr>
<td>GENERate</td>
<td>NORmalscores</td>
<td>NOTe</td>
<td>SCAtterplot</td>
</tr>
<tr>
<td>INPut</td>
<td>RANK</td>
<td>OPEN worksheet</td>
<td>STEm</td>
</tr>
<tr>
<td>INSert</td>
<td>RECode</td>
<td>OUTput</td>
<td>SYMbol</td>
</tr>
<tr>
<td>LOCK</td>
<td>SELECT</td>
<td>PARameter</td>
<td></td>
</tr>
<tr>
<td>NAME</td>
<td>SORt</td>
<td>QUIT</td>
<td></td>
</tr>
<tr>
<td>REAd</td>
<td>STATistics</td>
<td>RESTore</td>
<td></td>
</tr>
<tr>
<td>REMOVE</td>
<td></td>
<td>VDU</td>
<td></td>
</tr>
<tr>
<td>SHOW (?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNLock</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Simple Statistics

<table>
<thead>
<tr>
<th>Simple Statistics</th>
<th>ANOVA and Regression</th>
<th>Other statistics</th>
<th>Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALcule</td>
<td>ADD</td>
<td>FREquencies</td>
<td>COMmand</td>
</tr>
<tr>
<td>CHIsquare</td>
<td>ANOva</td>
<td>PERCentiles</td>
<td>ECHO</td>
</tr>
<tr>
<td>CIRCular</td>
<td>CORrelations</td>
<td>PRObabilitie</td>
<td>ERRor</td>
</tr>
<tr>
<td>DEScribe</td>
<td>DROp</td>
<td>ACF</td>
<td>MISssing</td>
</tr>
<tr>
<td>GAMma</td>
<td>ESTimates</td>
<td>CCF</td>
<td>WARn</td>
</tr>
<tr>
<td>GROup</td>
<td>FACtor</td>
<td>DIFFerences</td>
<td></td>
</tr>
<tr>
<td>NDF</td>
<td>FIT</td>
<td>LAG</td>
<td></td>
</tr>
<tr>
<td>POWER</td>
<td>INDicator</td>
<td>MOving average</td>
<td></td>
</tr>
<tr>
<td>RELative risk</td>
<td>INTERactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROW</td>
<td>IPF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIZE</td>
<td>ONEway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STATistics</td>
<td>REFIt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TABLE (old version)</td>
<td>REGression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TINterval</td>
<td>TERms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAriances</td>
<td>YVARIate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZINterval</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To find out more about any command, the Reference Guide is included in Instat’s help. If you type a command and press <F1>, you will go straight to the reference help on that command. Alternatively use Help ⇒ Reference Guide to give the screen as shown in Fig. A1.3b.

**Fig. A1.3b Help on the commands**

Choosing any letter in **Fig. A1.3b** takes you to a one-line summary of all the Instat commands that begin with that letter, as shown in **Fig. A1.3c**.

**Fig. A1.3c Finding help**

A third route to the help is illustrated in **Fig. A1.3d**. Here you start with the dialogue for the task. All dialogues have a Help button. At the top of the help screen that appears, is an option to look for the corresponding commands, as shown in **Fig. A1.3d**. That takes you to the help on that command.
A1.4 Using the occasional command

Even when we usually use the dialogues, we occasionally type a command into the “Commands and Output” window. This is mainly because it is simple to type and therefore so much quicker. Examples include the following

- **Removing columns to tidy the worksheet**
  
  : remove x8-x20, instead of **Manage ⇒ Remove (Clear)**

- **Using Instat as a simple calculator.** This is described in Chapter 8, section 8.4.
  
  : ? 2+3+4, instead of **Manage ⇒ Calculations**
  
  : x9=x8*2.5
  
  : ?mean(x6)

- **Defining a column as a factor.**
  
  : fac x5 4 instead of **Manage ⇒ Data ⇒ Define Factor**
  
  : fac X5 L1

- **Getting the last graph again.** Possibly with a small change.
  
  : replot instead of going back to the **Graphics ⇒ Plot** dialogue
  
  : replot ; href 0 adding a horizontal reference line
  
  : replot; xaxis 0 10 changing the x-axis

- **Checking what was done recently.** This is particularly useful when helping to find errors that beginners have made.
  
  : review Currently there is no dialogue for this. An example is given in **Fig. A1.4a.** This shows the last 16 commands (some resulting from using dialogues) that we used when preparing this Chapter. In **Fig. A1.4a** the “c” in front of the command indicates that Instat processed the command. If there is an error, then the “c” is omitted.
You can also use Edit ⇒ Recall ⇒ Last Command, or <Shift><F11> to bring the last command back. This is useful if it needs to be modified slightly, because it can be edited and is then re-executed when you press <Enter>.

### A1.5 Keeping a record of your work

Take any worksheet, with some data in x1 and x2. Now use Edit ⇒ Command Logging ⇒ Start, as shown in Fig. A1.5a.

Give it a name as shown in Fig. A1.5b. Then continue with your work. For illustration we suggest the following analysis.

- **Statistics ⇒ Data Summary ⇒ Describe.** Use x1 and x2, with the default statistics.
- **Statistics ⇒ Simple Models ⇒ Normal, One-Sample** for x1, changing the confidence level to 99%
- **Statistics ⇒ Regression ⇒ Simple,** with X1 as the response variable, and X2 as the explanatory variable. Include a plot of the model together with the fitted line.
Now use **Edit** ⇒ **Command Logging** ⇒ **View Log File**. The results will be something similar to that shown in *Fig. A1.5c*.

The log file has kept a record of what you have done. In addition it could be used to repeat the analysis. This might be on a different set of data, or perhaps you decide later to repeat the analysis, but just change a small part.

*Fig. A1.5c Edit ⇒ Log ⇒ View Log File*

You could choose to run the commands again directly from the log file, but we suggest otherwise. Instead, use **Edit** ⇒ **Edit Macro** ⇒ **New**. This gives an empty editing window. Now copy that part of the log file across that you want to use in the future. An example is shown in *Fig. A1.5d*. In the next Section we show how the resulting file can be used.

### A1.6 Using sets of commands

*Fig. A1.6a* shows an example of a set of commands for a complete analysis. We describe alternative ways that these commands can be entered and executed.

The first is to type in the “Commands and Output” window. Type everything in bold in *Fig. A1.6a*, i.e. both the commands and the data. Whenever you press <Enter> that command will be executed and Instat will return a prompt, when it is ready. Those are the parts in *Fig. A1.6a* not in bold.

**A1.6a A simple example**

<table>
<thead>
<tr>
<th>Commands</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| : CRE @TEST | **Introduction** Creates a blank worksheet  
| Title : Test example | Gives it a title – leave a blank line if there is no title |
| : ReAd X1 X2 X3 | **Data Entry** (Numbers may be separated by one or more spaces, or by commas) |
| data 1: 10 15 20 | |
| data 2: 12 14 19 | |
| data 3: 22 11 26 | |
| data 4: 41 9 22 | |
| data 5: | |
| : DisPlay X1-X3 | **Data Analysis** Display the data in the Output window |
| : INFo | Look at information about the worksheet |
| : Plot X1 X2 | A simple graph |
| : Describe X1 X3 | Display of summary statistics |
| : TiNterval X1;CON 99 | T-interval, with 99% confidence limits |
| : YVa X1 | |
| : REGress X2;Plot | Simple linear regression with a plot of the fitted line. |

This used to be the standard way of using Instat.
The second method is to type these same commands into an editing, or macro window. Possibly the data would first be entered directly into a worksheet. Then you just need the commands for the analysis. If you already have the commands for the analysis, as described in the previous section, then use **Edit** ⇒ **Edit Macro** ⇒ **Open**. You could add the **CReate** and **REAd** commands to that file, to make it similar to **Fig. A1.6a**, or just run the analysis.

**Fig. A1.6b** File of commands

<table>
<thead>
<tr>
<th><strong>Editing [Unnamed]</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>DISPlay X1-X3</td>
</tr>
<tr>
<td>INFO</td>
</tr>
<tr>
<td>PLOT X1 X2</td>
</tr>
<tr>
<td>DEScribe X1 X3</td>
</tr>
<tr>
<td>TINterval X1; CON 99</td>
</tr>
<tr>
<td>VYA X1</td>
</tr>
<tr>
<td>REGress X2; PLOt</td>
</tr>
</tbody>
</table>

If you do not have the commands from a log file, then use **Edit Macro** ⇒ **New** and type the commands, or paste them from **Fig. A1.6a**. An example is in **Fig. A1.6b**.

Then, with the editing window active, use **Submit** ⇒ **Current Window** as shown in **Fig. A1.6c**.

The file with the commands can then be saved and used again on another occasion, or with another set of data.

To show a third possibility, first save the data in the current worksheet, into an ASCII file. The **File** ⇒ **Import/Export** ⇒ **Output** dialogue is shown in **Fig. A1.6c** and saves the data into a file called test.out.

**Fig. A1.6d** File ⇒ Export ⇒ Output

<table>
<thead>
<tr>
<th><strong>Output of data</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data to be output</td>
</tr>
<tr>
<td>X1X3</td>
</tr>
<tr>
<td>Data columns:</td>
</tr>
<tr>
<td>X1</td>
</tr>
<tr>
<td>X2</td>
</tr>
<tr>
<td>X3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Output formats</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple ASCII (.txt)</td>
</tr>
<tr>
<td>Comma delimited (.csv)</td>
</tr>
<tr>
<td>Basic with line numbers (.bas)</td>
</tr>
<tr>
<td>Genstat (.gn)</td>
</tr>
<tr>
<td>SPSS (.sp)</td>
</tr>
<tr>
<td>SAS (.sas)</td>
</tr>
</tbody>
</table>

Then the commands in **Fig. A1.6e** input the data from a file and do the analysis. The important line is

**Fig. A1.6e** Using ASCII data file

```
restore: warn off
core @test
test worksheet for Appendix 1
read x1-x3; file @test.out
join x1-x3; into x4; fac x5
display x1-x5; wid 6
info
plot x4 x5
box x4;by x5
stats x4;by x5; nea; sde
yva x4
oneway x5
```
In Fig. A1.6e we have also changed the commands, to illustrate more of the Instat language. This could now be used to process any set of data, just by changing the name of the file.

The file of commands, shown in Fig. A1.6e, is supplied with Instat, and called Example1.ins. This is sometimes called a “macro”. In the version that we have supplied, we have added a further feature that we describe in the next section, and its use is shown in Fig. A1.6f. Instead of having to edit the macro, whenever we want to analyse a new set of data, the user is asked for the name.

Fig. A1.6f Making the macro more friendly

In the examples in this section we have shown how we can move from typing a few commands to the development of a system for automating the analysis of any set of data. We continue in the next section, by showing how macros can be run.

A1.7 Using a macro

A macro is just a set of Instat commands, such as the examples we saw in the last section. You can use macros without having to write them yourself. The “trick” is to get someone else to write them! We supply a little library of macros with Instat and others can be added locally. Here we show the different ways that you could use a macro, once you or someone else has written it.

The macros that we supply all have the extension “ins”. The first way to use a macro is to start with Edit ⇒ Edit Macro ⇒ Open. This gives the dialogue shown in Fig. A1.7a and provides access to the macros you may have written, or been given. They may be in your current working directory, or they may be in one of the libraries.

Fig. A1.7a Edit ⇒ Edit macro ⇒ Open

Once you have chosen the macro it opens into an editing window. You can then examine and edit it, if you wish. To run it, make it the active Window and then use the Submit menu as shown in Fig. A1.7c. Alternatively, as shown by the options in the menu in Fig. A1.7c, you could just run a single line, or a part of the macro file that you have selected.
The second way of using a macro is the “Run Macro” option, shown in Fig. A1.7d. This gives a similar dialogue to Fig. A1.7a. Ignore the option about "parameters". The macro runs immediately that you click OK.

A third method is useful if you plan to use the same macro repeatedly. Use the Submit ⇒ Add Macro to Menu option, which gives the dialogue shown in Fig. A1.7e.

We give the macro an informative name, which is “Example from Appendix 1”, in Fig. A1.7e. Pressing OK appears to do nothing. But if you then use the Submit menu again, as shown in Fig. A1.7f, the macro has been added to the bottom of the menu. In future you now just have to click on that option to run the macro.

The final way of running a macro is to call it as a command, as shown in Fig. A1.7g. You simply type the name of the command file (macro) with an @ in front, so Instat knows it is a macro. A slight variation is shown in Fig. A1.6h. Here you type

: exec @example1

instead of just @example1. The macro runs in the same way, but in Fig. A1.7h the commands within the macro are also shown.
If you are “into commands” then the fact that you can call a macro as a single command implies that you could include a macro within another macro. Then each macro could be simple, but together they could become powerful.

There is one restriction on the names of a macro. In the olden days file names were restricted to 8 letters. This still applies either if you call the macro in a command as described in Fig. A1.7g or A1.7h, or if the macro is called by Instat itself, i.e. by another macro. If you call it from a dialogue, then the front-end of Instat automatically shortens the name before passing it as a command. To be on the safe side use short names for macros.

A1.8 Adding Help to a macro

Once you have a few files with commands it is useful to keep information about what they do, and perhaps why they were written. We have provided a simple system to encourage you to keep this information.

Start by using Edit ⇒ Edit Text ⇒ New. Now type some text into the editor as shown in Fig. A1.8a. Save what you have typed, using the same name as the macro, but the extension “aid”. When you use File ⇒ Save As you will find that this is a one of the suggested file types.

In this case, as the macro is called Example1.ins, the help file is called Example1.aid. Alternatively you could prepare this file in a word processor, or any other editor, as long as you save it in simple text or as an “RTF” file.
Now, use either Edit ⇒ Edit macro ⇒ Open, or Submit ⇒ Run Macro and choose the name of the macro, in this case Example1, as shown in Fig. A1.8b. If Instat finds an associated help file, then the “Help on Macro” button is enabled. Use this button to show the help file, as in Fig. A1.8c.

This scheme of providing your own help is just for the macros that are in your current directory, or in your local library. If you would like to provide your own help on the macros that we have provided, then make a copy of them in one of these directories first.

A1.9 Writing a macro

We have already seen examples of simple macros in this chapter. All you need to do, as shown in Section A1.5, is to use a log file to record the commands that are generated as you use Instat. Then these commands can be used again, possibly in a modified form. We call this a macro.

That is our starting point in this section. We use an example from Chapter 19, Section 19.4, to illustrate how macros can be made into more general tools is called rowreg, because it calculated the slopes of regression lines from each row of the data.

The data from Section 19.4 are shown again in Fig. A1.9a. The are from 9 bee-hives, where the area has been measured on 5 successive occasions.

We already have the results of using the macro in X13 of Fig. A1.9a. The macro has calculated the slopes of the regression lines we needed and put them into the column. We
show how they have been calculated, by following the steps again for the first row of data, i.e. to produce the value 70.4 that is in X13 in Fig. A1.9a. This will help us to decide how to produce the macro we need.

**Fig. A1.9b** Manage ⇒ Remove

**Fig. A1.9c** Manage ⇒ Transformations ⇒ Select

We first remove any columns of data from x14 onwards, as shown in **Fig. A1.9b**. Then, as shown in **Fig. A1.9c** we use the Manage ⇒ Transformations ⇒ Select dialogue to transpose part of the first row of data into a column. We use x15, so that x14 is available for the final results. For reasons explained in Chapter 19 (that are unimportant for the ideas of writing a macro) we use the data in x4-x7, that correspond to the 2\textsuperscript{nd} to the 5\textsuperscript{th} time points.

Then we put the values 1 to 4 into x16. We can now use the Statistics ⇒ Regression ⇒ Simple dialogue to give the slope of the regression line. The result is in **Fig. A1.9d**, and we see from the equation of the fitted line that the slope is 70.4. That is the first value we want. The data, in x15 and x16 that we used for the regression, are shown in **Fig. A1.9e**.

**Fig. A1.9c** Results, showing slope of 70.4

**Fig. A1.9d** Calculation of slope

```plaintext
\begin{verbatim}
\begin{verbatim}
SELECT X4:X7;INTO X15;IF row=1;TRA
Number of cases = 1

Simple Linear Regression

\textit{\texttt{R\texttt{\char'15}X15:Reg X16}}

\begin{verbatim}
DETAILS OF THE FITTED LINE
\end{verbatim}
Fitted equation : X15 = 838 + 70.4 * X16
Standard error of slope : 16.4 with 2 d.f.
95\% confidence interval for slope -0.16449
R-squared : 0.902
\end{verbatim}
\texttt{x17=(x15-mean(x15))*x16-mean(x16)}
\texttt{x14(1)=sum(x17)/case(x16)}
\end{verbatim}
```

But this dialogue is not suitable for a macro, because it has no facility to store the results back into the worksheet. We would have to do a lot of copying and pasting, or we might have to type the values into the worksheet.

The full regression dialogues would be better, because they have the facility to save the estimates, but a simpler alternative is shown at the bottom of Fig. A1.9d. Mead, Curnow and Hasted (1993), page 164, give the formula for the slope of a regression line as

\[ \text{Slope} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \]

This, in Instat, is the bottom two lines in Fig. A1.9d and has provided the value of the slope in x14 in Fig A1.9d. We see that one calculation is correct, by comparing this result with the output in Fig. A1.9c.

In Fig. A1.9f we have used these commands as the basis for a simple macro that calculates the slopes of the first two hives.

The macro in Fig. A1.9e is reasonable, but it is lucky that there are only 9 hives. If there were 100, then the macro, written in this way, would be very tedious to construct.

We need a “looping” facility and this is shown in Fig. A1.9f. The new feature that we have used is

```
Loop ;repeat 9
<commands>
```

We have also introduced a new integer variable, called “%1”. This is initially set to take the value 1. In Fig A1.9f we then have the line

```
%1 = %1 + 1
```

to increment the value. That is how we are able to fill the column x9 successively with the slopes.

The final macro that is in Fig A1.9f is a reasonable solution to the particular problem that we had. It is also easy to change, if we were to alter some aspects of the problem. For example, if there were 50 hives, then we just change the “repeat 9” to “repeat 50”.

We have shown here that macros to solve particular problems are easy to write. You could quickly add a help file, as described in the last section, and build a small personal library of useful macros for yourself.

You might then decide to turn this macro into a simple general tool. However there is a difference between solving a particular problem and providing a general tool. As a general tool we would not want users to have to alter the macro every time their problem changed.
To illustrate the difference we have extended the macro from that shown in Fig. A1.9g and supplied the more general version in the library. It is called “rowreg” and gives the slopes of a set of regression lines for data that is in rows, rather than columns. You can examine that macro, if you wish, by using Edit ⇒ Edit Macro ⇒ Open. Or you can run it, as described earlier in this chapter. If you run it, then you will have to be prepared for 2 questions, as shown in Fig A1.9g and A1.9h.

**Fig. A1.9g  Running rowreg macro**

**Fig. A1.9h  A second prompt for input**

Further information of the macro facilities in Instat are provided in Chapter 14 of the climatic guide.

A1.10 In conclusion

In this chapter we have progressed from the recording of the Instat commands, using a log file, to the use of simple macros, to conduct routine analyses efficiently. The last example was an extra “row-regression” facility that was not possible in Instat without a macro.

Everyone used to start using computers, by learning a language. Earlier this was Basic, or Fortran and later it might be Pascal or C. Or with statistics packages it would be the command language of SPSS or SAS. Now users start by learning Windows, and Word or Excel. The standard ideas of variables, or loops or branching with “if statements” is taught much less.

We hope that the gentle introduction to Instat’s commands will help users who move to other statistics packages, because the same arguments apply, that we have used in this chapter. The standard statistics packages all have a much more powerful language component than Instat.

Excel comes with the powerful Visual Basic for Applications (VBA) language included. The power of this language is shown by the wide range of add-ins that is provided, including our own SSC-Stat.

Among the standard statistics packages the type of language they include is related to the history of the package. Those, like SPSS, SAS and Genstat, that started long ago, have a more procedural type of language, while S-Plus and R are built on the more modern S-language.
References


CAST (see http://www.cast.massey.ac.nz).


EPIINFO (see http://www.cdc.gov/epiinfo/)

Excel (see http://www.microsoft.com/office/excel/)


GENSTAT (see http://www.nag.co.uk/)

GLIM (see http://www.nag.co.uk/)


MINITAB (see [http://www.minitab.com](http://www.minitab.com))

MLWIN (see [http://www.ioe.ac.uk/multilevel/](http://www.ioe.ac.uk/multilevel/))


Netica (see [http://www.norsys.com](http://www.norsys.com)).

nQuery [http://www.statsol.ie/nquery/nquery.htm](http://www.statsol.ie/nquery/nquery.htm)

ORIANA (see [http://www.kovcomp.co.uk/](http://www.kovcomp.co.uk/))

PAINT SHOP PRO (see [http://www.jasc.com](http://www.jasc.com))

pass [http://www.ncss.com/pass.html](http://www.ncss.com/pass.html)


R (see [http://www.r-project.org](http://www.r-project.org))


SAS (see [http://www.sas.com/software/](http://www.sas.com/software/))

SDR (Species Diversity and Richness – see [http://www.pisces_conservation.com](http://www.pisces_conservation.com)).


S-PLUS (see [http://www.mathsoft.com/](http://www.mathsoft.com/))

SPSS (see [http://www.spss.com/](http://www.spss.com/))

SSC-Stat (see [http://www.rdg.ac.uk/ssc/](http://www.rdg.ac.uk/ssc/))

STATA (see [http://www.stata.com/](http://www.stata.com/))

STATISTICA (see [http://www.statsoftinc.com/](http://www.statsoftinc.com/))

STELLA (see [http://www.hps_inc.com/](http://www.hps_inc.com/))


WINBUGS "*Exploratory Data Analysis*. Duxbury Press, Wadsworth Inc." (see http://www.mrc-bsu.cam.ac.uk/bugs/)
INDEX

Durbin-Watson statistic, 226

E
Ecological study, 15, 24
Entering the data, 65
Experimental data, 50, 112, 139
Exponential distribution, 125, 170, 171, 277
Exporting, 79
Extreme value, 170

F
F, 19, 20, 41, 66, 170, 182, 184, 186, 191, 203, 217, 218, 219, 226, 227
Factor columns, 19, 20, 61, 66, 67, 69, 130, 160, 222, 287, 288
Factor levels, 74, 163, 205, 206
Factorial treatment structure, 196, 199
Format, 60, 69, 99, 100, 147, 148, 162, 223
Formulae, 134
Frequency tables, 139, 142, 148, 159
Friedman test, 245

G
Gamma, 170
General linear models, 220
Generalised linear models, 112, 210, 211, 212, 234, 294, 295

H
Help in Instat, 3, 8, 45
Help on statistics, 48, 49
Help on teaching, 50, 51
Histogram, 16, 127, 272, 277

I
Importing data, 58, 67, 72, 73
Index numbers, 81, 87
Indices, 23, 25, 44, 55, 81, 84–91, 93, 106, 107
Installation, 7, 12, 13, 39
Interaction columns, 196, 198, 222, 251
Interactions, 168, 197, 200, 204, 210, 251
Interactive language, 297
J
Jack-knife, 91

K
Keeping a record, 303
Kruskal-Wallis test, 245
KURtosis, 134

L
Label columns, 19, 61, 66, 288
Lack of balance, 290
Linear regression, 183, 184, 211, 212, 213, 214, 215, 218, 220, 222, 223, 224, 226, 228, 235, 304
Lock/unlock, 76
Loess fitting, 234
Log file, 303, 304, 305, 309, 312
Logical calculation, 83
Log-linear models, 10, 50, 211, 239, 253, 294, 295
Looping, 311

M
Macros, 10, 40, 71, 72, 107, 294, 300, 306, 308, 309, 311, 312
Mann-Whitney, 245
MEDian, 134
Menus and dialogues, 10, 27, 28, 36, 47
Missing values, 79, 80, 160
Mixture, 168, 169, 211, 218, 220, 233, 293, 298
Monitoring data, 15, 21, 87, 139, 281
Moving average, 260, 262, 264
Mstat, 109
Multilevel models, 294
Multiple dichotomy, 154
Multiple levels, 22, 95, 102, 107, 112, 140, 148, 157, 158, 160, 210, 267, 274
Multiple range tests, 205
Multiple regression, 218, 219, 222
Multiple response, 154, 155

N
Netica, 294
New worksheet, 29, 57, 59, 64, 78, 101, 163, 277, 284
Non-linear models, 235
Non-normal models, 180
Non-orthogonal, 206, 207, 208, 209
Non-parametric methods, 10, 234
Normal distribution, 112, 114, 123, 124, 125, 168, 175, 178, 186, 212, 234, 241, 275, 277, 294

O
ODBC, 1, 68, 73, 74, 75, 76, 80, 107
One-way, 189, 190, 206, 229
One-way ANOVA, 190, 245, 249, 267
One-way tables, 10, 136
Open from library, 58
Organising the data, 9, 55, 110
Oriana, 274
Orthogonal, 189, 193, 194, 195, 197, 202, 204
Orthogonal polynomials, 205

P
Parallel lines, 216, 217
Pattern, 144, 176, 184, 189, 190, 195, 207, 211, 252, 261, 263, 267, 280, 281
Percentage points, 133, 149, 151, 171, 172, 177, 178
Percentile, 133, 134, 153
Permutation, 286
Poisson, 32, 170, 180, 207
Polynomial, 205, 214, 229, 230, 231, 232, 234
Polynomial regression models, 229
Poverty index, 23, 85, 86
Presenting tables, 145
PRESS statistic, 225
Probability models, 167
Probability plot, 123, 186, 223, 275
PROC Mixed, 274
Profiles, 267, 268
Programming, 294, 295
Pseudo-random numbers, 275

Q
Qualitative data, 10, 50, 187, 210, 237, 239, 240, 241, 243
Quantitative treatments, 204

R
R, 111, 312
Rainfall index, 262
Random samples, 10, 48, 59, 60, 275, 278, 281–91
Rank sum, 246
Raw data, 22, 105, 113, 121, 139, 163, 178, 181, 240, 241, 242
Recoding, 82
Reference Guide, 1, 48, 90, 299, 301
Index

Registration, 11
Regression diagnostics, 211, 214, 222, 223
Regression models, 211
REML, 274
Repeated measure, 255
Resize Worksheet, 57, 63, 79, 96
Rice survey, 51, 52, 82, 127, 140, 162,
165, 180, 190, 227, 238, 240, 245, 289
Richness, 89, 90, 91
Row operations, 269
Row Statistics, 81, 83, 84, 85, 87, 89

S
Sample size, 169, 178, 275, 281, 287,
289, 291
Sampling distribution, 169, 170, 276
Sampling scheme, 165, 287, 289
Sampling with replacement, 286
SAS, 6, 107, 111, 165, 234, 274, 312
Scatter-plots, 119
Selecting and Stacking, 98
Shannon, 90
Significance test, 112, 175, 179, 180, 183,
184, 256
Simple models, 10, 30, 32, 137, 175, 176,
181, 192, 208, 209, 244, 248, 303
Simple regression, 182
Simulation, 10, 52, 275, 277, 279, 280
Simulation games, 52
SKEwness, 134
Spline models, 234
Split plot experiment, 199, 207
S-Plus, 111, 130, 234, 274, 312
Spreadsheet, 5, 6, 18, 20, 27, 39, 43, 62,
65, 67, 75, 79, 80, 95, 107, 110, 111,
113, 128, 138, 158, 160, 256, 273, 277
SPSS, 5, 6, 44, 53, 54, 80, 109, 111, 157,
164, 165, 253, 312
SSC-Stat, 39, 40, 41, 43, 80, 107, 115,
128, 138, 186, 312
Stack, 70, 95, 96, 99, 245, 246
Standard deviation, 134
Standard error, 168, 172, 175, 179, 180,
182, 183, 191, 203, 208, 224, 225
Standardised residuals, 225, 226
Standardising, 83
Stata, 111, 165, 234, 274
Statistica, 52, 111
Stem and leaf plot, 127
Stepping-stone, 5, 6, 43, 111

Strategy, 6, 13, 28, 107, 109, 110, 111,
113, 115
Studentised residuals, 225, 227
Subsets, 81, 96, 100, 107, 233
Sufficient statistics, 178
Summary tables, 142
Survey data, 49, 112, 139
Survival studies, 10, 255, 273
Systat, 111

T
t, 170, 172, 177, 179, 184, 225, 227, 245,
266
Tables, 30, 113, 128, 139–65, 240, 247,
250, 251, 253, 267, 289
Tables of percentages, 249, 252
Teaching percentages, 148
Temperature data, 263
Text editor, 6, 37, 70, 76
Time series, 10, 21, 130, 255, 259, 260,
261, 264, 267, 273, 274
Time to events, 256
Transformations, 30, 32, 81, 82, 207
Transpose, 100, 101, 310
Trellis-plots, 130
Trend, 129, 130, 176, 260, 261, 262, 264,
265
Trimmed mean, 53
Tutorial Guide, 8, 9
Two-level, 158, 267

U
Unstack, 96

V
Von Mises distribution, 273

W
Web, 7, 11, 44, 50, 51, 52, 93, 112, 116,
161
Weibull, 170
Weighted least squares, 233
Weights in tables, 151
Wilcoxon, 245
Winbugs, 277, 294
Without replacement, 285, 286, 287
Worksheet editor, 59, 65, 78
Worksheet font, 35, 62
Worksheet information, 63, 76
Worksheet library, 59