R-Instat

Table of contents

Introduction

Welcome to R-Instat. We hope that you will agree with us that R is a wonderful system. We hope you will also agree that R-Instat makes a few of the resources in R really easy to use. In this Section we first describe why we have developed R-Instat.

We then consider what might be new to you. Of course "new" will depend on what you have been using so far. If you have been using a spreadsheet for statistical work, then "new" could be that this is your first statistics package. If you have already started to use R, then this way of using R may be what's new.

We then look briefly at what we consider to be the "special" features of R as used via R-Instat, going menu by menu. Then we look at who R-Instat is intended to support.

The Case for R-Instat

The case for R-Instat was made in our crowd-sourcing campaign. This started in July and finished on 8th November 2015.

The link to the campaign site is here. It starts with a 3.5 minute video which is also on You Tube.

We claim that the first stage in the campaign does seem to be working. We have been developing the software and you are using the first beta version. The development has been an excellent partnership with the work largely being done in Kenya and Tanzania, plus in Reading at the SSC (Statistical Services Centre), now also with SSD (Statistics for Sustainable Development). So it really is software developed in Africa and for Africa, but we think it could be useful much more widely than that.

We have basically been constructing an easy-to-use front-end to R. It is designed to be as easy to use as any statistics package can be. So it can be a first statistics package for users who are reasonably comfortable with a spreadsheet. It can be used by anyone, but we explain its target audiences in the section titled "Who is R-Instat for?".

Usually realism sets in once the programming starts. We have tried to be realistic, but the quality and variety of R packages is just astounding. So we have become steadily more excited by what is possible. The phrase "standing on the shoulders of giants" is appropriate. The vision, dedication and quality of work of the group who started R and now those who continue to add packages is incredible. We hope we can be successful, both in broadening the groups of people who can use R and also in providing an environment that can prompt improvements in the teaching of statistics.

Tell us what you think!

What's new

Easy to use

Our main aim has been to make R-Instat easy to use. By that we mean it provides a way to use the statistical system, called R, easily. So almost everything can be done from typical "Windows-style" menus and dialogues.

We have written a "front-end" in VB.Net. You interact with this front-end. When you press OK, VB then sends the corresponding commands to R. Then R sends the results back to the Output window in the front end. The two main Windows show the data and the output. There are other Windows, described here.

New "spreadsheet"

There is a comprehensive spreadsheet system for viewing and editing data. When you open an R-Instat data frame, you will get a spreadsheet view of the data.
You may have many data frames open at a time. The tabs at the bottom of the sheets (data frames) are displayed in the same way as in spreadsheet packages. (We use software called Reogrid, see https://reogrid.net/ and have been very pleased with both its facilities and the responsiveness of the developers to questions concerning their software.)

New graphics

We have always been impressed with the ideas in the book called the *Grammar of Graphics* by Leyland Wilkinson. Hence we have uses the R implementation of this system, called ggplot2, by Hadley Wickham, wherever possible in R-Instat. We therefore think that R-Instat's graphics (well really Hadley Wickham's graphics) are now great.

This graphics system also typifies why R-Instat is becoming an interesting tool to help you with learning or practising statistics. It is not because we are wonderful (though of course we are!). It is because of the wonderful people who conceived of the S and then the R system for statistics. And then the literally 1000's of talented people who have contributed packages to R. We are simply providing you with an easy-to-use introduction to some of these packages. The graphics is an example, where the only criticism we hear of the ggplot2 package is that it can be difficult to learn to use effectively.

New statistical tools

We make use of these R statistics packages to provide comprehensive facilities in R-Instat to organise your data and then for your analyses. We are still not aiming to be "comprehensive". R-Instat is still designed partly as a "stepping-stone". It assumes you are comfortable using a spreadsheet to look at data. If R-Instat is then your first statistics package, then we hope you will find it surprisingly easy to use, and with lots that is familiar from your experience with a spreadsheet.

If you ever need more, then we hope R-Instat has prepared you well to either migrate to using R itself, (probably through RSudio) or to use any of the standard statistics packages of your choice. And, from our R-Instat "stepping stone" you should again find either route to be easy. So, while you add statistics packages, you remain able to concentrate on analysing your data, rather than being diverted (unless you wish to) into having to spend much time learning the new package.

New Ideas

One reason for developing R-Instat was to contribute to improved teaching and learning in statistics. Over the past 40 years we have continually been impressed with the way that developments in statistical methodology have made the subject simpler. Hence it should be both easier to teach and also easier for students to digest. This is both for people who are specialists in statistics and for those who specialise elsewhere (e.g. biology, agriculture, health, sociology) but need some statistical skills to handle their data.

We have equally continually been disappointed that these simplifications in the subject seem to be largely absent in many courses in statistics. A common complaint by students (and employers) remains that courses are too theoretical and hence do not give students a picture of how to use statistics or how to process their data. In addition, most courses are "tool based". They teach particular methods, and often use simple data sets for which there is a single right answer. The real world is "problem based". Real problems often require a variety of methods that each illuminate different aspects, and there is rarely a single "right" answer.

We hope the structure of R-Instat can tempt some trainers to consider adding problem-based courses, and also to take advantage of the new unity in the subject. We used to have a Graphics menu and a Statistics menu, like most other menu-driven statistics packages. We have changed that into a ***Describe*** and a ***Model*** menu. The Describe menu (largely tables and graphs) is to tempt trainers and users to emphasise descriptive statistics more, because it is so important in analysing real (and often large and messy) data sets.

And we have organised the modelling menu in a different way. Most training courses still seem to teach topics in their order of mathematical complexity. So, for example, simple and multiple regression, for normally distributed data, is taught before generalised linear models. We wonder if this still need be the case, given that the student doesn't do the mathematics of model-fitting in the same way now.

Instead, we have organised R-Instat's modelling menu in the order of how many columns of data are involved in fitting the model. For example estimating the parameters of a single normal distribution is just one column, and so is the estimation of the parameters of any other distribution. Simple linear regression has two columns, and so does logistic regression, or even comparing two samples, or one-way ANOVA. And so on!

This order of the menus ties more simply with our common strategy when giving statistical support. We find consultees often arrive asking for the "best test" to use for their data. This is, for us, based on the misconception that statistics consists largely of significance tests. So, if only we (as professional statisticians) could tell them the appropriate test to use, then they could simply apply it, their analyses would be over and they could continue with topics they enjoy (and understand) more.

Instead we ask (seemingly irrelevant and annoying questions) on what their objectives were? Why did they collect the data? If their objectives corresponded to a single "column" or variable, then this is the column to be analysed. More often the objectives involve more than one column, or at least there are further columns that help to explain the variability in their data. So then we can start by looking at these columns of data - initially descripively, and then perhaps via a modelling approach.

And also …

We have taken advantage of the fact that R is object orientated. So we keep the data in an ***Instat object***. Roughly an ***Instat object*** is a collection of R data frames, (sheets) together with the associated meta-data. We are not trying to build a database system, but we think it is useful to keep information on how columns of data relate to each other, and similarly how data in different sheets are related. For example this is what a spreadsheet package does when it generates pivot tables - and they compete well with the tabulation facilities in most statistics packages.

And graphs are also R objects and so are (regression) models, etc. So those objects can also become part of an Instat object too!

This is all quite exciting!

Help

Still to be updated for R-Instat

The help has changed completely. You can now access the complete user guides, particularly the Introductory and Climatic Guides from the Help menu. There is also help on each dialogue. This help includes information on where to find further information about the method of analysis and which sample datasets you can use to try the analysis.

The R-Instat Introductory Guide (318 pages) has not yet been completely rewritten. It is designed to support the use of R-Instat and to assist users who wish to improve the ways they teach and learn statistics. It is also designed to help users to decide on their software strategy. As such, it describes alternative software for tasks that cannot be handled fully within R-Instat.

The R-Instat Climatic Guide (300 pages) has also not yet been completely rewritten. Both these guides are available as pdf files, for those who wish to produce a printed copy.

Changes Menu by Menu

To be updated for R-Instat

In this section we add to the list of changes, described earlier and summarise some of the innovations we have made in this implementation. We consider menu by menu.

File menu

Edit menu

Manage menu - now called Organise

Describe menu (old Graphics menu, plus descriptive statistics)

This is perhaps our tour-de force. The implementation of the Grammar of Graphics in R through ggplot2 is a way to produce wonderful graphs. We have worked hard to provide as much of ggplot2's flexibility as possible.

Model menu

Climatic Menu

A special menu for analysing climatic data.

To be updated from the old Instat

The main change is the existence of easy-to-use dialogues for all the common climatic analyses. In addition the making Instat+ easier to use, they also provide a new way to access to all the information in the Instat+ Climatic Guide.
The changes, which include a "calendar" option on some dialogues, make analyses simpler for sites where the start of the year is not the 1st January.
Most of the menus include options to plot the results. The analysis of (daily) data is usually a 2-step process. The dialogues simplify the first step, which is to get the summary values for each year. Two dialogues that are in Version 2 are called Examine and Process. They are designed to ease the second stage, which is often to calculate risks or return periods.
Some of the dialogues on the other menus, (besides climatic) have been specially enhanced with climatic analyses in mind. Examples include, fitting gamma models, in the Statistics => Simple models menu and

fitting of circular data in the Statistics => Summary menu. In the Manage menu the Transformations => Select and Stack dialogue is particularly intended for climatic summaries of the daily data.
Many of the macros in Instat+ were particularly to help with climatic analyses. Hence the new macro system should provide even better access for analysts. You can even arrange for a macro that is frequently used to be added to the Windows menu system.
We are also particularly interested to see if the dialogues for the Markov chain modelling facilities, which were added to Instat in 1997, are now accessible to more potential users.

Who is R-Instat for?

We start with spreadsheet users, who may never have used a statistics package before. But users who have data to analyse and are prepared to consider using a simple statistics package, We expect you to find much that is familiar, compared to using a spreadsheet for your data organisation and analysis. We also hope you will be pleasantly surprised how simple it is to add a statistics package.

And you don't need to give up using a spreadsheet either. Through the R package called rio, which is "behind the scenes" in R-Instat we import very well from Excel and other packages too, and can also export back to Excel.

We have four particular types of user in mind:

1. Students of statistics (and their lecturers). You may be intending to be a statistician, or studying statistics to support your main subject area.
2. Researchers, particularly in agricultural research who have their data to analyse.
3. Staff who need to analyse climatic data. Here we consider particularly staff from National Meteorological Services, but also those in many other organisations for whome climate variability and change is an issue.
4. Staff in National Statistics Offices who have survey data to process.

In our campaign we concentrated primarily on the first group, i.e. on students of statistics. We have also kept to the idea that the software should primarily be developed in Africa, and this was largely in Kenya and Tanzania. The team continues the work as we aim towards our first release in July 2017.

Acknowledgements

The original Instat started over 30 years ago and was on a BBC microcomputer. Later, four staff members were the main contributors to the Windows upgrade, namely Ian Dale, Colin Grayer, Joan Knock and Roger Stern, from the SSC, University of Reading. For the Windows version we would particularly like to continue to acknowledge the support of the UK Met Office, without whose funding that transformation would not have been possible. Their particular interest was the climatic facilities in Instat and now in R-Instat.

We also acknowledge the support of the Statistical Services Centre, University of Reading, that enabled more of the general statistical and data manipulation facilities to be included in the release of Instat+ and also the Department of Applied Statistics for their encouragement to include their statistical "games" within the Instat environment.

Our Statistical Services Centre once were the Biometric Advisors to the DFID Natural resources Division. As part of our proactive work for them, we produced the "good-practice" guides for DFID research projects. We are grateful for their encouragement to distribute this information freely, more widely than to DFID project teams. Their inclusion in Instat and now in R-Instat is an experiment on a further way of distribution. These

guides became a book called "Good Statistical Practice in Natural Resources Research" published by CABI. We are very grateful that CABI have now given permission for a pdf version of this book to be distributed with this software.

Many people and groups have contributed to R-Instat. In particular we acknowledge, with many thanks, the contributions, together with the implied vote of confidence, from ther 426 people who contributed to our crowd-sourcing campaign. The main partners have continued to support the work. For example we were able to combine training in AIMS Tanzania with development work, while based in their wonderful centre in Bagomoyo. In the UK the new SSD (Stati8stics for Sustainable Development) has now grown from the original SSC and has continued to support the development work.

The current programming staff continue to be based at AMI in Western Kenya and SAMI continues to support, both AMI and the ADI work.

The funds from the crowd-sourcing campaign have now been supplemented by funds from CCAFS to facilitate the transformation of R-Instat into a successor to Instat for climatic analyses, and we are most grateful for this support. More recently we now have funds from the Global Challenges Research Fund to adapt R-Instat to be able to analyse publically available data on corruption.

Background to R-Instat

The pages here describe the background to R-Instat. How did it start and grow, and how was it transformed from Instat into R-Instat. We also try to answer some questions you may have on R-Instat

Instat before R

Instat from the start

Instat was written by statisticians engaged in teaching, research and consultation, in collaboration with programmers. It began in 1983 as a series of programs for the BBC Microcomputer for teaching statistics, written by Bob Burn. He later worked as a Principal Statistician for the Statistical Services Centre following many years working as a consultant in Mauritius.

Instat was first used on a training course of statistics in agriculture given in Sri Lanka at the end of 1983. It was then further developed and marketed from mid-1985.

In 1987 Instat became available commercially for the PC compatible range of machines. The first shareware version (V6.1) was released in 1994, V6.2 released in September 1996 contained many new general features, while V6.5 had new chapters for the Instat Climatic Guide and new commands to model rainfall data.

Instat was used in the UK and abroad by a wide range of companies, research institutes, schools, colleges, universities, and by private individuals working at home. It has been used in many courses on general statistics and on statistics in agriculture, health and climatology courses, given in Reading, Sri Lanka, Africa, China and elsewhere.

During this period, development of the package continued in collaboration with staff from The University of Reading under the guidance of Roger Stern, who returned to work at Reading, after seven years at ICRISAT in Niger in the 1990s. During Roger's absence, Joan Knock ran the Instat office at Reading.

Instat in Windows

The first test release of Instat for Windows was in 2000, in time for it to be tested and then used on 10 week training courses in Nairobi in May 2000 and Reading in July 2000. The first public trial version was released in October 2000, with further releases in Summer 2001 and September 2002.

These versions were produced by Roger Stern, Colin Grayer, Ian Dale and Joan Knock, all staff at the Statistical Services Centre, The University of Reading. Funding for the first release under Windows was provided by the UK Met. Office at Bracknell and the Statistical Services Centre. The Statistical Services Centre funded the development of the 2002 release.

Early versions of Instat were totally in BBCBASIC(86), with some sections in Assembler - thanks to Richard Russell (who was primarily responsible for BBCBASIC on the PC). The Windows version added a front end, written in Visual Basic.

Since 2003 successive versions of Instat were in Windows, using versions of BBCBasic for Windows (http://www.bbcbasic.co.uk/index.html) for the Instat server, with the front end remaining in Visual Basic.

For the past 10 years the main users of Instat have been those who wished to analyse climatic data.

We bow to the inevitable in moving Instat from BBCBasic to R, but those who remember the BBC microcomuter and its BBC Basic language will appreciate the nostalgia that some of us feel, particularly as BBC Basic continues to be available.

R-Instat

R-Instat is a front-end to R. It builds on many of the ideas behind the original Instat, though none of the code from the original software has been transferred. It hardly has a history!

R-Instat was conceived in early 2015, and the development was then made possible through the success of the crowd-sourcing campaign in mid-2015. Development started in the Autumn of 2015, and the first beta relaease of R-Inst5at was on 8 November 2016. Many readers will know of this date as the occasion of the US election competion between Hilary Clinton and Donald Trump. We will remember it as the one year anniversary of the end of our crowd-sourcing campaign and hence as the data of the release of the first beta version of R-Instat.

Why did or didn't you...?

Some common issues have been raised, when we started to crowd-sourcing campaign, and as we have been developing R-Instat. We list them here together with our answers.

1. Why did you bother? There are already so many statistics packages?
2. Why did you use R? Why not just promote the use of Stata (or another common or free package)?
3. Why didn't you use an existing "front-end" to R?
4. Why didn't you start by engaging more with R developers?
5. Why didn't you start by consulting more with R-users, to find what they needed?
6. Why did you insist on developing most of the code in Africa?
7. Why did you use data frames rather than data tables?
8. Why haven't you optimised your code better?
9. Why did you use VB.NET for the front-end?
10. Why did you only develop for Windows computers?

What's special about R-Instat?

Learning and Teaching Statistics

Applied Statistics at the University of Reading has a long history of innovation in the teaching of statistics. We give some information here and the R-Instat Introductory Guide will also include some of our ideas on the teaching of statistics.

In Africa, and perhaps also elsewhere, Universities and other organisations do not always have an easy-to-use statistics package that can be freely distributed to their students. Lecturers sometimes cite this lack to explain why they are still teaching in traditional ways - ways that over-emphasise formulae, and that students often claim are too theoretical. They also over-emphasise the methods and formulae that were already available in the 1960s with courses largely following the order of mathematical complexity. This is no longer needed, and courses can now concentrate more on ideas and concepts. In R-Instat we have tried to show what this can mean in practice.

In addition, a greater use of realistic sized data sets, could tempt some courses to spend far longer on aspects concerned with organising and describing data (graphs and tables). These are a large part of the practise of statistics. They are also fun and are sometimes largely absent in training courses.

When the Statistical Services Centre were the Biometric Advisors to the DFID Natural resources Division we produced a series of "good-practice" guides. We are grateful for DFID's encouragement to distribute this information freely.

A series of Case Studies, to show the varied roles of a statistics package in teaching was added to Instat as long ago as 1990, when a Teaching Pack was produced.

This menu collects ideas and resources concerning the use of the computer to support the teaching of statistics. Most of the resources are also freely available outside Instat for trainers who wish to make use of them while continuing with a different statistics package.

The motivation and contents of this menu are described in more detail under the Strategy option.

The Case Studies provides a set of examples of ways in which the computer, and a package, such as R-Instat can be used in teaching. These case studies can be for reading, though the descriptions have yet to be updated for the Windows version.

All the datasets are also provided so students or teachers can evaluate the ideas.

Finally we list other resources that may give ideas for trainers and trainees in relation to their statistics teaching. These are partly from Reading, but mainly from other sites that we have found useful.

Strategy

We have collected together some of our ideas on the teaching of statistics. Our main theme is that the integration of computers and the relevant software can improve greatly the teaching and learning 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. But we do need to improve our teaching, particularly to users who are not intending to become statistics specialists. Statistics is widely taught, but generally remains unpopular. We pay a heavy price for this.

Access to a computer can help in a variety of ways.
1. It enables a statistics package, such as Instat+, to be used within the course.
2. Special programs can be used to illustrate particular concepts, or to allow students to simulate a data generating process.
3. Documents and other reference materials can be made available on-line. The Help => Teaching => Resources submenu describes some that we feel are particularly useful for the sort of training courses where R-Instat may be used.
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 R-Instat dialogue help.

Here is an example of the teaching ideas about Balanced ANOVA, which can be obtained from the dialogue help file or from Statistics => Analysis of Variance => Orthogonal dialogue. Note that these Teaching Ideas also link to several other help items, relevant to ANOVA.

Most dialogue help has examples of datasets that are supplied with Instat+. For example the Reptile dataset that is used in the Introductory Guide to illustrate the analysis of a factorial design.

There is an index of all the Datasets, Teaching Ideas and Further Information on the dialogues.

Changes in teaching

Many people attend statistics courses, or try to teach themselves some statistics. After standard courses some still feel ill equipped to handle real data. The use of the computer within the teaching process can help to broaden the syllabus of a statistics course and make it more relevant.

Specific points are as follows:

It is easier to illustrate design and planning issues

For example the simulation games, such as the rice survey or the chick experiment, can be used to illustrate different sampling schemes or experimental designs. This is no substitute for involving students in a real survey, but it can cover concepts of planning in a way that is otherwise difficult given the time constraints of a standard training course. We say more on the games under Using other software.

More time can be devoted to descriptive statistics

Without a computer, many courses just teach how particular techniques are used on relatively trivial (and uninteresting) data sets. Thus they do not spend much time discussing when to use them.

With access to a computer some of the data sets can be larger and more interesting. They can have more structure and hence be more realistic and challenging. More time can then be devoted to data exploration, and to alternative methods of display and summary of the data.

Learning how to summarise data construction is both useful and simple. Hence students can become proficient, even if they later find inference topics, such as significance testing, to be difficult.

Some data management can be taught

Most real sets of data are quite large and their analysis proceeds in stages. Early stages usually involve selection and summary. The resulting summaries then often become the raw data for a subsequent stage in the analysis.
Without a computer, students are often given only trivially small data sets, hence they do not see this data management stage. If students have no experience of these simple early stages in data handling, it is hardly surprising that they find it difficult to cope with large data sets later.
Without a computer, many data sets are semi-processed before being used. Sometimes courses, and text-book examples, start with such data (grouped frequencies, counts from chi-square tests etc.) without giving any indication of the raw data. This can be confusing. Case Study 11 illustrates the problem.

The teaching of standard topics can be illustrated

Within R-Instat, the Statistics ⇒ Simple Models dialogues can help the teaching of statistical inference for 1 and 2 sample problems. Statistics ⇒ Simple Models ⇒ Probability Distribution provides plots of the distributions and illustrates concepts such as the central limit theorem. Similarly the Statistics ⇒ Regression ⇒ Simple was constructed to help the teaching of simple regression.
We find the ideas of statistical inference are often poorly understood. Perhaps our good-practice booklet on statistical inference can help.

The use of statistics packages can be taught

Statistics packages are used routinely in the real world, hence it is useful to encourage their use within a course or to analyse project data. The way Instat+ is used is very similar to many other packages. Hence users who have to move to "heavyweight" statistical packages later, should find the transition is easy because they are already familiar with the concepts.

Putting the ideas into practice

The Introductory Guide describes these ideas in more detail and examples of how these ideas translate into practice are shown in the Case Studies. Above all, we have found that access to a statistics package, such as R-Instat helps our courses to be more enjoyable to teach - and we hope, to receive.

Not with R-Instat

Our ideas for improving the teaching of statistics are relatively independent of the package that is used. They do not need R-Instat. So, if you already use a statistics package, then you might already be doing all that we suggest, or can easily adapt the ideas to the package that you are using.

For example, many Universities and Colleges use the statistics package called Minitab. This also encourages good statistical practice. Sites may wish to use the statistical games and the Good Practice Guidelines, but not complicate their courses by introducing another statistics package. In this case, simply download the games and guidelines from www.reading.ac.uk/ssc/.

Most students have some experience of using spreadsheets. Some may wish to continue with use of, for example Microsoft Excel for their statistical work.

Spreadsheets can be a very useful tool, particularly for simple data manipulation and descriptive statistics. Most users appreciate their flexibility, but this also makes them very easy to misuse. They must be used with "discipline" if the aim is to encourage "good statistics".

Our Good Practice Guidelines include one on the use of a spreadsheet for data entry, on when to move from a spreadsheet to a database package, and one that looks at strengths and weaknesses of Excel for statistics. We also have an "add-in" to encourage good statistical practice with Excel, see our web site www.reading.ac.uk/ssc/ for more details.

Other statistical software

Using other software in addition to a statistics package

Teaching Concepts

There are other useful packages and special programs to teach particular concepts. In the menu Help => Teaching => Resources, we have collected some Web addresses that have resources related to the teaching of statistics. Many programmes have been written to illustrate different topics, such as the shape of distributions, the central limit theorem, or the idea of confidence limits.
We claim that R-Instat complements the use of these special programs. R-Instat is designed to help resource persons and students to be able to illustrate many concepts, within a familiar (statistical package) environment. If you want to emphasise a particular concept, then almost certainly a special program will be better. Students would have to be shown how to use this different program, but this is usually a simple process, if it is in a familiar Windows or spreadsheet type of environment.

Statistical Games

At Reading we have used "statistical games" for many years, to support our teaching. These may be played on the computer or the computer may be used to generate the materials, for "games" that are then played "by hand".
Most of the games allow students to practice collecting their own data in a variety of ways. This enables aspects of the design or planning phase of a survey or experiment to be taught. The experiment or survey then produces simulated data that have to be analysed. This enables the whole process, from planning, to analysis and reporting to be taught.
Statisticians are keen to recommend that clients should come for statistical advice, before the data are collected. But remarkably little of the training of statistics in conventional courses is devoted to real experimental and survey design. What is taught instead is largely the analysis phase, but using data from different designs.
It is then not surprising that scientists are weak at the statistical ideas related to the design phase and also that statisticians, following a conventional training course, are ill-equipped to advise on practical design. Without such simulation programs it is quite difficult to teach the concepts of design within the time constraints of a training course. They are no substitute for real practical experience of the planning phase, but they do help in teaching some of the key concepts. Most are based on real case studies. The use of case study materials is valuable as well.

Good Practice Guidelines

Encouraging good statistical practice is much broader than simply the use of a statistics package. We have included a set of Good Practice Guides that cover the following areas:

- Overview of Topics
- Planning statistical investigations
- Data entry and management
- Data analysis
- Presentation of results
- Combining qualitative and quantitative information

Each of these guides is deliberately short, in the hope that users, who might be put off by a long document, will be prepared to read them. We have also produced the guides in different formats (written booklets, HELP files and pdf files) to cater for different user preferences.
In aspects that involves the use of a statistics package we propose that good statistical practice includes the following ideas:

1. Users should be encouraged to start with the raw data, rather than semi-processed data. The computer can easily be used for the initial summary steps and users still have access to the raw data for checking purposes or for alternative summaries.
   R-Instat therefore provides data summary facilities that enable initial summaries to be calculated and stored in further columns for the next stage of the analysis.

2. The "exploration" of the data is to be encouraged. Instat+'s facilities for plotting are designed to help with this. Tabular data can also be presented usefully, including coding options so that important features can be highlighted.

3. Interactive analyses are to be encouraged, i.e. the results of one stage of the analysis may suggest the next actions. In this regard Instat+ tries to give adequate output that does not overwhelm the user.If more output is needed, then the options can be specified when the dialogue is used. The describe dialogue provides an example.
   Some statistics packages give voluminous output, and this is not helpful for users, particularly when they are learning the subject.
   The multiple regression dialogue epitomises the role of interactive analyses. It allows for the regression model to be modified by adding and dropping variables.

4. Results from analyses (residuals and fitted values) should always be easy to save. They can be used for the display of results or in checking on the validity of the model used.

Resources

Guides (pdf files) that can be viewed or printed
The following three guides provide support for the type of training course where Instat may be used:

1. Confidence and Significance: Key Concepts of Inferential Statistics (pdf version)
2. The Statistical Background to ANOVA (pdf version)
3. Moving from Chi-Square Tests to Log-Linear Models (pdf version)

The three guides mentioned above are also available as Help files.

The Inference Guide is one of our "Good-Practice Guides" and provides a brief and non-mathematical review of the key concepts of statistical inference, such as confidence limits and significance tests. This is the basic area that many students find difficult. Once these concepts are understood, much of statistics falls into place.

Instat also has many dialogues in the statistics menu that try to help the teaching of these ideas, for example Probability Distributions, Simple Models, Regression => Simple Linear and Analysis of Variance => OneWay.
Analysis of Variance is another subject that is often poorly understood.
This guide (The Statistical Background to ANOVA) was originally produced for special software, which is now, with the permission of Professor Pearce, and asru, incorporated into Instat. So this guide, together with the associated datasets and the use of Statistics => Analysis of Variance => General is partly designed to support the teaching of ANOVA.

Generalized linear models is a most important development in statistical methods. The context is described in our Good Practice Guide on Modern Methods of Analysis. Within this class of models, the use of chi-square tests can now be extended into log-linear models. The special guide called "Moving from chi-square tests to log-linear models" describes this approach. The method can be illustrated using the Statistics => Regression => Log-Linear Models dialogue.

Other resources at Reading University
Our web page, www.rdg.ac.uk/ssc includes further resources. In particular we have an add-in to support good statistical practice with Excel.
We also have training notes for a variety of courses. For example there are detailed notes for a course on the analysis of experimental data, together with a series of datasets. These have been prepared in collaboration with ICRAF and have been used on various courses in Benin, Guinea, Kenya and Tanzania.
What is available at other web sites.

Here we are not trying to be comprehensive in what we provide. Instead we have looked for information, and particularly web sites that are good in their own right and also ones that can lead to further resources. It is a "web" after all.
If you go to only one site, then we suggest the Computers in Teaching (statistics) web site at http://www.ltsn.gla.ac.uk/ at Glasgow. The CTI has ceased, but this site currently still provides lots of information relating to teaching in statistics at further education level. It includes a list of statistical journals. CTI has been incorporated into the new "Learning and Teaching Support Network Centre for Maths, Stats and OR". It includes a database of teaching and learning resources that are freely available over the web. This database is divided into the following categories:

- Textual Resources
- Interactive Routines
- Multimedia
- Software
- Virtual Stimulation Objects
- On-Line Course
- Computational Aids

- Reference Tools
- Assessment
- Teaching Articles

Also look at the Royal Statistical Society (RSS) Centre for Statistical Education, at http://science.ntu.ac.uk/rsscse Its remit is broader than the CTI and its successor above, being interested in education at all levels.
Its aim is stated to be "To promote the improvement of statistical education, training and understanding at all ages". It specifies school and higher education, both for specialists in statistics and those who need some statistical skills. It also provides education for users who need statistics as part of their professional development. Finally for "society as a whole".

The RSS Centre includes information from the journal Teaching Statistics, an international journal for teachers of pupils between 9 and 19.

In the USA, try http://www.amstat.org/publications/jse for information on the Journal of Statistical Education.

What can be downloaded?
The sites above both give extensive information on statistical software, including many reviews.

If you found that Instat was not sufficiently powerful for your needs, then the R-system will provide all the power you need. It is free and available on a wide range of systems. If you need a very powerful statistical system, but you also require the ease of use that Instat provides, then you will have to look also at some of the commercial packages.

Case Studies

In these Case Studies we have tried to show some of the ways in which the use of a computer can change the way statistics is taught. The Case Studies are on a wide variety of topics, and the titles are fairly self explanatory.

Please note that these files have not yet been updated for the new Windows versions. That is on our wish list. But we thought they are still useful, even if they look a bit "old-fashioned"

.

Case Study 1: Introduction to the use of R-Instat
Case Study 2: Simulating sampling distributions
Case Study 3: Testing goodness of fit
Case Study 4: Squashing data
Case Study 5: Survey data
Case Study 6: The t-test - Bookwork plus
Case Study 7: Using simulation to avoid messy algebra
Case Study 8: Hot data
Case Study 9: Least squares without calculus
Case Study 10: Throw away your statistical tables
Case Study 11: To group or not to group

Most statistics courses taught without access to computers are fairly boring, particularly if the audience is not statistics specialists. We hope these case studies indicate ways in which courses can teach more, and at the same time become more fun.

You may be sceptical about the use of the computer in teaching for many different reasons. We accept that the examples shown here could be done with other statistics packages, we accept that there are other types of software that could also be used in statistics courses. We also accept that over-dependence on the computer is undesirable.
But after all this, if you currently teach statistics without much use of the computer, or have not changed your teaching much, since computers were introduced, look at these examples with an open mind. We would be surprised if there was nothing that was useful in its own right, or at least, in firing your imagination to produce a further study that would be of use

.

The command files to run these case studies are called CASE1 to CASE11 and can be found using Edit ⇒ Edit Macro ⇒ Open and selecting a case study from the Library.

Teaching ideas related to dialogues

This section provides ideas for teaching that are associated with particular dialogues. This has yet to be updated from the "old" Instat.

One-way ANOVA

The One Way ANOVA is important in teaching. It can be used to explain the components of an ANOVA table. And unlike the ANOVA table for simple linear regression, the F-test has a different role to the corresponding t-tests for comparing means. It is therefore a good example to explain the role of significance testing.

The ANOVA plots are illustrated and described in Section 16.2 of the Introductory Guide. They are used to illustrate the different components of an Analysis of Variance table. This is usually used in conjunction with similar plots from simple linear regression models to emphasise that ANOVA is a useful general technique.

Now produce as a single figure with the 4 plots

ANOVA

Students (and researchers!) often find ANOVA to be a confusing subject. They may produce the ANOVA table as an automatic reflex, but then ignore it in the analysis and just use the treatment means.

The Analysis of experimental data guide discusses this issue.

Central Limit Theorem

Use the Statistics ⇒ Simple Models ⇒ Probability Distributions dialogue, with the plotting option, to look at the sampling distribution of the mean. Use the Manage ⇒ Data ⇒ Random Samples dialogue to look at histograms of empirical sampling distributions.

For example, with the Probability Distributions dialogue, sample for proportions (Binomial sampling) or from the Poisson distribution. Use a small sample size and look at the exact sampling distributions for the proportion or mean, together with the fitted normal. See how the approximation is improved if the sample size (or the mean of the Poisson) is increased.

You can show the same pattern when sampling from the exponential distribution, i.e. the exact sampling distribution of the mean gets progressively closer to the normal distribution as the sample size is increased.

Then use the Manage ⇒ Data ⇒ Random Samples dialogue with at least 20 samples and choose to plot the summaries (mean or median) of each sample.

Chi-square Test

In Case Study 11 - "To group or not to group" we emphasise that statistics teaching should normally start by showing students the raw data. The chi-square test is used for semi-processed data, where the counts are given. The step that produces the counts, which in Instat is the Statistics => Tables dialogue can be used to produce the data for this dialogue.

The chi-square test is limited to counts for 2-way tables. This can now be generalised to multiway tables. This is discussed in the Instat Teaching Guide on log-linear models, see also the Statistics => Regression => Log-linear Models dialogue. This approach is more satisfactory, because it incorporates the logic of models for counts within the general regression-modelling framework.

Circular Statistics

This is one of the types of data mentioned in the Introductory Guide, Chapter 19.5 to question how our teaching of statistics can change once we have simple facilities for processing data. Do we still always need to understand the theory, before a new topic is introduced?

Further examples of circular data are also in the Climatic Guide Chapter 12.3.5.

Column Statistics

In our teaching we emphasise that the "raw data" should be entered into the computer, since summary statistics are more reliably calculated using a spreadsheet or statistics package than "by hand" or with a calculator. In addition, storing the raw data makes checking and archiving easier. The Column Statistics facility in R-Instat simplifies the task of deriving summary statistics, and is a key to our assertion that the raw data should be computerised.

The most common complication in data analysis is the collection of information at different levels. For example, a survey may give information at village, household and person level; or an ecologist may record some information at a quadrat level and then record details of ten plants in each quadrat. The Column Statistics dialogue permits information to be moved up a level, e.g. from person to household level, or from plant to quadrat level. See the Introductory Guide Chapter 9.6 for more examples.

The fitted and deviation features are useful in teaching the ideas of analysis of variance. An example of their use in summarising time series is in Section 19.3 of the Introductory guide.

Confidence Limits

Many of the dialogues, particularly those for fitting simple models, i.e. Statistics => Simple Models => Normal, One Sample, etc, provide for plots that include confidence limits. These plots are partly to show the data, but also to help users understand what is meant by a confidence interval. If the mean is being estimated, then they are an interval that usually includes the true mean.

We find that students are often not clear on the meaning of a confidence interval. The graphs usually also show that a confidence interval for the mean is NOT the same as an interval that includes most of the data. This is the same confusion as the difference between the standard error (of an estimate) and the standard deviation (of a set of data).
To clarify this concept the dialogues for the Statistics => Simple Models = Normal, One Sample and Normal, Two Samples also include the option to add the "prediction interval", i.e. the interval that is likely to include an individual point (rather than the mean).

The Simple Models =>Normal, One Sample dialogue has two further options. One is to give simple summary limits. These are simple descriptive statistics and are xbar +/- 1.96 s, for the 95% limits. These are close to the prediction limits, but the latter include the uncertainty in the estimation of the population mean by the sample mean.
The second option is to show the confidence limits for specified percentage points. This shows there are other statistics you may wish to estimate, besides the mean. Whatever you estimate should be accompanied by a measure of precision, i.e. the standard error, or a confidence interval.

Correlation

The use and misuse of correlations is described in the Climatic Guide Section 11.5. The points are general, even though the illustrations are concerned with climatic data.

Crop Index

Chapter 10 in the Climatic Guide describes the use of this method of analysis and Chapter 12.2 describes the irrigation modelling scheme.

As a simple water balance budget, this method can provide a useful summary of the rainfall data that is relevant to a particular length of season. It should not be thought of as a crop model, but the way in which any crop model can utilise many years of climatic data is shown through this facility.

Similarly, the way in which the success of an irrigation scheme can be measured through its effect on a crop model is shown by the irrigate option and the methodology is independent of the particular irrigation or crop model that is utilised.

Daily Climatic Data

Part of our case for users to analyse daily climatic records, rather than simply monthly or other summaries is that the daily data can easily be summarised when needed. The Climatic ⇒ Summary dialogue shows how easy this step has become.

Following this dialogue the data are often processed further to give percentage points of the summarised data. How this is done in a simple way, for rainfall, is described in Chapter 5 of the Climatic Guide. Users can, if they wish, fit gamma models, and a case study in the climatic guide compares 5 different methods of processing the summarised data.

Data Structure

The appropriate analysis depends on the structure of the data. This is an important idea and we consider 3 examples, all of which can be illustrated.

1. Simple or stratified random sampling? Try the Statistics => Probability Distributions dialogue, normal samples and look at a mixture of 2 distributions. If you choose to sample from the mixture then the spread of the data may be much larger than if you are able to stratify the population and sample separately from each part.

2. Paired or independent samples? Both the Simple Models => Normal, Two Samples and Proportions, Two Samples allow you to indicate whether the samples are paired. When they are paired it is appropriate to look at the differences, and a failure to recognise this element of the structure often leads to a much less precise analysis than should be done. This same concept can be illustrated using the Statistics => Probability Distributions dialogue, with the plot option. Choose to look at the distribution of the difference between 2 normals. When there is effective pairing this corresponds to a high correlation between the 2 random variables. Use examples to show how much this correlation reduces the standard deviation of the difference.

3. Structure in regression models. The concepts of structure with 2 samples extends to more complicated modelling. Users often forget that their data might have well-defined groups when they use regression models. The structure of the data will usually lead to models that include the idea that there are different groups of data. This is just the idea of including factors in regression models. The Regression => Simple with groups dialogue shows the situation with a single factor, while the general situation leads to what is called a General Linear Model (GLM).

Display Daily Data

It is important that trainees are encouraged to look at the daily data, before proceeding with the analyses. We used to print out the tabular results that can be obtained from this dialogue for the 11 years in the

samsmall (or other worksheets with daily data) file.

This is easily done from this dialogue and can be viewed in the output window.

Evapotranspiration

The Climatic Guide, Chapter 9, gives a review of the scope of agricultural climatology and then introduces the reasons for using Penman's formula in calculating evapotranspiration. The case for Penman-Monteith is also described in detail.

Factor Columns

To understand what is done with this dialogue, have the spreadsheet open when using this dialogue. It will jump to show the column that was created. Then use the Climatic => Summary dialogue to summarise the daily data. There you can give monthly (or decade etc.) totals for rainfall data, or means for temperatures, etc.

Frequency Tables

Survey data are usually analysed by giving summary tables as is shown in the Case Study 5 - "Survey data". This facility is also useful to teach the overused subject of percentages. For a 2-way table it shows that many percentages are possible and the user has to be clear on what is meant by 100%.

Generating Random Samples

The ability to generate random samples is often used in teaching, as is shown in Case Studies 2 -"Simulating sampling distributions" and 7 - "Using simulation to avoid messy algebra".

For users interested in climatic analyses the climatic dialogues illustrate the use of simulation to model rainfall data.

Indicator (dummy) Variables

It is important to consider regression modelling as a general idea, where the model can include variates and factors. The use of indicator (dummy) variables helps in teaching how factors can be included.

Reword from here
However, the fact that Instat's regression facilities do not cope automatically with factors shows its limitations.
More powerful packages, such as Genstat, and SAS, (with PROC GLM), allow regression models that contain a mixture of factors and variates and do all the "indicator" type of calculations automatically.
This is similar to the General Linear Models dialogue in Instat, but Instat still stores the indicator columns in the worksheet. Hence, with Instat, the size of a regression problem is limited by the number of columns, and with factors in the model, it is easy to need 100 columns.
On the other hand, the visibility of the indicator columns, both in the worksheet, and in the General Linear Models dialogue is a powerful teaching tool. An example is in the Introductory Guide, Chapter 17.6.

Log-Linear Models

This dialogue illustrates the advances that have been made in statistical methods recently (i.e. in the last 30 years!), specifically in the flexible ways that data from non-normal distributions can be analysed within a modelling framework.

The dialogue shows that the simple chi-square test is a very limited way of analysing contingency tables. A separate guide is provided that includes examples and compares the chi-square test and this type of modelling in more detail.

Mixture of 2 Distributions

This is a plotting option from the Statistics => Probability Distributions dialogue. Uses include the following:

a)

It is primarily to encourage users to consider the structure of their data. For example, if the data are heights of people, then the structure normally includes the fact that the sample includes both sexes. It is usually better to use stratified sampling, than just to consider the data as from a single population.

b)

You can add the title to the plot and it then gives the standard deviation of the mixture. Use this to discuss how much larger a simple random sample needs to be to give the same precision as a stratified sample.

c)

The plots show that a wide range of unimodal and bimodal distributions can arise as mixtures.

d)

If you have no option but to sample from the mixture, then the inference problems are not trivial. For example, with the Normal model there are up to 5 parameters to estimate (mean and sd of each distribution and the proportion of the data from one distribution.) These are fewer if you know the proportion, or can assume equal spread for the 2 distributions. The estimation of the parameters is too complicated for Instat, but is simple in some statistics packages, such as Genstat.

Plots of the difference between 2 normal distributions

From the Statistics => Probability Distributions dialogue one option when plotting normal distributions, is to investigate the distribution of the difference between two normal variates.

Start by assuming independence. Then the difference has variance equal to the sum of the two variances. This is often taught, but not particularly intuitive. Perhaps seeing the graph can help. If you add the title, it gives the mean and standard deviation of the difference.

The same result can be seen when analysing data using Statistics => Simple Models => Normal, Two Samples, where the confidence interval for the difference is wider than the individual confidence intervals.

A second point from the Probability Distributions plot arises when the variates are correlated. With a positive correlation the variance of the difference is decreased. This is related to the value of using paired samples when comparing two treatments in the Two Samples dialogue with Normal models.

Multiple Levels

Data at multiple levels arises frequently in practical data collection exercises. For example,

- Person within house within postal code

- Child within class within school
- plant within quadrat
- leaf within plant within pot
- replicate within person within treatment group

Simple statistics courses teaches students the importance of replication, but with multilevel data we must clarify at which level do we have the replication.

It is important to introduce students to hierarchical data for two reasons. The first is statistical, and particularly the topic of replication, mentioned above. The second is because of the data management issues that are introduced.

In simple statistics, we have our data within a single rectangle, but data at 2 levels, implies 2 rectangles, and so on. There are important practical issues of data management that should become part of our training in statistics. For example, if data entry is into a spreadsheet, then we should normally use 2 sheets, with one for each level. Our good practice guides give more information.

This Manage => Reshape => Merge dialogue, shows how data can be copied down to a lower level. The converse is to summarise up a level and this uses the Statistics => Summary => Column Statistics or the Statistics => Tables dialogues. (If you are using a spreadsheet, then Pivot tables would do the same thing.)

As we explain in our good practice guides, multiple level data provides some motivation for using a database package, such as Access or EpiInfo for the data entry and management. Then most statistics packages provide an ODBC link, so they can be used for the analysis.

Such issues have, in the past, not been considered as part of training in statistics. We believe that their omission is a major reason for people finding that real-world statistics (even school-level practical projects) is SO different to what they have been taught. We challenge teachers of statistics to at least say where their students are being taught practical issues of data management.

Non-Parametric

Non-parametric tests are important, but they are overused. There is still the assumption by some scientists, that data should be analysed using a non-parametric test, if they are not from a normal distribution. This can divert the analysis from other methods that are now available for the analysis of non-normal data.

We describe the analysis of qualitative data in the Introductory Guide, Chapter 18.4. The good-practice guide, titled Confidence and Significance looks briefly at non-parametric methods in Section 7 and discusses a general modelling framework in Section 9.

Normal Models 1

To be copied over

Normal Models 2

To be copied over

Orthogonal Polynomials

To be copied over

Poisson - one sample

This model has been added, partly in its own right, but also to be able to introduce topics concerned with the modelling of non-normal data that go beyond the capabilities of Instat

Poisson models are for another example of a different type of data, namely a count. The help on the Proportion, One Sample dialogue considers types of data more generally.

In teaching, such data are often presented as grouped frequencies. We describe, in Case Study 11, why we think this is often confusing for students. This dialogue allows for the data to be given in various forms to reinforce this point. The raw data can be given, or the frequencies, or just the summary statistics, i.e. the number of observations and the mean.

Probability Distributions

The top of the Statistics => Probability Distributions dialogue is as follows:

The Plot option is designed to support the teaching of many concepts associated with the use of simple probability models. These include the following:

a) Plots of the density functions (Normal, gamma, t, …) or the individual probabilities (Binomial, Poisson).

b) Plots overlaid with the normal distribution, to assess the normal approximation. (This is usually the Normal distribution with the same mean and standard deviation. For the t-distributions it is the standard normal, N(0,1).)

c) Plots may also show the sampling distribution of the mean, when sampling from the Normal, exponential, Gamma, Bernoulli or Poisson models. The user can choose whether to plot the exact sampling distribution, or the normal approximation, or both. This illustrates the central-limit theorem.

d) A title can be included with the plot. This is used to indicate a key property of the plot, like the standard deviation of the sampling distribution.

e) Plots can show 2 distributions, to motivate the teaching of 2-sample statistics.

f) Plots with 2 distributions can also show a mixture. This can be used to discuss the value of stratified, rather than simple random sampling.

g) There are special plots to look at the distribution of the difference between 2 Normal distributions.

Proportions

One sample

One important point is for users to consider the different types of data they might collect. The data could be

categorical and the simplest case is where there are just 2 categories. If there are more than 2 categories, then they may be ordered, say from VERY GOOD to VERY BAD, or unordered, like the reasons readers give for liking Harry Potter books.
There are now statistical methods in more powerful packages for handling what is called multinomial data. Instat provides some ideas for the simplest case, namely binomial data. You can always construct binomial data by splitting any more-complicated columns into 2 categories. So an important part of this dialogue is for users to decide on their definition of "Success".
Statistics becomes simpler when users can always visualise the raw data. In teaching the analysis of proportions this point is often hidden, and the user starts with just the number of successes, r, out of the n trials. The binomial dialogues provide for this by allowing the data to be given, either in their raw form, or as summary statistics.
Chapter 18 of the Introductory Guide is concerned with the analysis of qualitative data and this dialogue is used for illustration in Section 18.3, where we discuss teaching ideas and Section 18.4 which is on non-parametric methods..

Two samples

The teaching points under Proportions, One Sample apply equally here.
A further point concerns the structure of the data. It is important for users to see that the design of their study leads to a well-defined structure of the data and the analysis depends on this structure.
Here this is illustrated with the PAIRED option. The data are paired if you get a YES/NO answer from the SAME people on 2 occasions. In terms of structure this is the same as the paired t-test for continuous data (see the Normal, Two Sample dialogue). How the analysis changes is perhaps less important than the idea that it should change, and the user should always consider the structure before deciding how to summaries the data.
In this case the analysis for paired samples usually leads to considering the change-overs - that is the equivalent of looking at differences for the paired normal case.

Random Samples

Check - this may be deleted
In teaching it is often convenient to generate data from a particular distribution and then process the data. With the Random Samples dialogue, the plotting options are mainly to support the teaching of simple statistical inference.

Data

This plots the individual values, either by observation (i.e. in the order they were generated), or by sample. This shows the sort of data that arises. The means of each sample are also indicated on the plot.

Summary

This usually needs at least 20 samples for an effective plot. As summary statistic, you may choose to calculate the mean, median, and/or standard deviation of each sample. Each summary statistic is then plotted as a histogram, with a title that includes the mean and standard deviation of the sample statistic.
When sampling from a non-normal distribution (try the exponential for example) the resulting plot is an illustration of the central limit theorem.
One use of these plots, when sampling from a Normal distribution, is to show that the sample mean is a more efficient estimator of the population mean than the sample median, i.e. it has smaller standard error. This is compensation for the lack of robustness to outliers of the sample mean, compared to the median.

Confidence limits

This is currently only available when sampling from a Normal distribution. The plot gives the specified confidence interval (default 95%) for each sample. It reinforces the idea that most of these intervals should include the true value.
Choose to sample with or without the variance being known, to show that when known, the length of the confidence interval is always the same. When not known the intervals are of different lengths, depending on the value of the sample variance.
In practice you only have one sample, but there are occasions when you know roughly the spread and feel that it is underestimated in your sample. What does this do to the interpretation of that particular confidence interval?

Regression, Simple

There are 4 alternative types of plot that can be given, that each illustrate a different point in teaching:

1. Plots of the data and the fitted curve. This can be considered as simple descriptive statistics. A useful basic plot to discuss whether a straight line is a sensible model.
2. Plot with confidence and prediction limits. This continues the discussion of these concepts that was covered when considering the Statistics => Simple Models for Normal models with one or two samples.
3. Plots to introduce ANOVA. This is the partner to the plots produced from the ANOVA => OneWay dialogue. It is to introduce the use of ANOVA as a general method for choosing between alternative models.
4. Plots to explain least squares. Users are challenged to fit models "by eye". They may use least squares or other methods that are more robust for the fitting. The best they can do for least squares can then be compared with the results from fitting the model.

This last option also shows how Instat would compare with the use of special software for illustrating this point. Any special software would be more dynamic, and allow the user to move the line interactively. Such software should be used if this concept is to be explained in detail. If, however, it is a passing point, then it may not be worth the time in installing and giving instruction in a new piece of software.

Regression with Groups

The main concept is that regression models can include factors as well as variates within the fitted model. Often data divide into groups, i.e. they correspond to different levels of a factor, and this might be important in the model that is fitted. The general idea is that the structure of the data should always be considered when developing a model.

The concept is illustrated here with models that are limited to a single variate plus one factor. The General Linear Model (GLM) dialogue is for more complex models.

It is usually simplest to use the graphical displays to explain how to choose between the different models. This explanation assumes that users have some idea of an ANOVA table to compare models. This is explained with the graphs associated with the Regression => Simple and ANOVA => OneWay, and should be used prior to the explanations associated with these models.

You may choose to fit estimates that are relative or absolute. The Introductory Guide Chapter 17.4 gives a full explanation with an example. Briefly the relative estimates give the intercepts and slopes relative to those of the last level of the factor. With absolute estimates you get the same estimated values that would result from fitting a simple line to each level separately.

Behind the scenes, the dialogue generates the indicator variables for the factor, before fitting the regression models. If you choose not to remove the generated columns after the fit, you can see what has been done.

Regression - lm

There are three points that are important from a teaching perspective:

1. The dialogue distinguishes clearly between the use of factors and variates in the fitted model. It also introduces the idea of interactions, both between factors, and between a variate and a factor.

It is probably useful if a discussion of such points is preceded by the use of the Regression => Simple with Groups dialogue, because this allows models with one variate and one factor.

2

The way the models are fitted can easily be illustrated by using the details checkbox. This shows that for any factor, or interaction, in the model there is behind-the-scenes work in generating intermediate columns that are then used in the fitting. This is the reason why the fitted model is not as easy to interpret as models that just include variates.

3

For the factors in the model the fitting options checkbox allows different choices in omitting the first or last level of any factors that are included in the model.

Spell Lengths

Redo or delete

One use of the simple results is to show how the full command works. The spell lengths are presented, on a daily basis, as shown in Fig. 7.1, for each year specified. They are either in tabular form or as a line graph, (Fig. 7.5) or both.

The tabulation of the full data is also useful as an exploratory tool, to study the structure of the data. For example, marking the date of the start of the season, together with a knowledge of the length of growing season for a proposed crop, is useful in a study of the frequency of dry-spell problems during each year.

In teaching we often use the 11 years of data in the @samsmall worksheet. This set of data is also useful when we have to explain why the longest spell length in April, say, is often longer than 30 days!

Start of the Rains

The simple results may be used to teach how the full command works or to help a user to choose a sensible definition of the start. It presents the rainfall totals, on a daily basis, for each year specified. They are either in tabular form or as a needle graph, or both. These data may be studied to see how the results would change when you alter the definition. Often we find that the date of the start is insensitive to the precise definition.

In teaching we often use the 11 years of data in the @samsmall worksheet. The results may be printed or displayed in the output window using 2 or 3 day totals. Students are then asked to note the starting dates "by hand", for 2 different earliest possible dates, say 1st April and 1st May. Some of these results are the same as the data that was used in the tutorial (link). This "hand analysis" helps students to understand what is done in the full use of the dialogue. It also is a useful moment to explain that this 3-stage analysis (look at the daily data, pick a useful summary value, analyse the summary) is the same for monthly totals or for dates of the start of the rains.

Summary Tables

Case Study 5 - "Survey data" and the Introductory Guide, Chapter 13 show how tables are used in the analysis of Survey data. It is useful for students to see that tables can easily be constructed that contain summaries as well as counts or percentages.

Tally

Case Study 11 - "To group or not to group" makes the case that courses should always show students the raw data. Textbooks often start the topic of chi-square tests, etc., with the summaries that are ready for the required test. This is confusing and the Group dialogue demonstrates how easy it is to move from the raw data to the required summaries.

Climatic Analyses

Since 1986 special modules were included for the analysis of climatic data. These commands were written by Roger Stern, Hugh Hack and Joan Knock. Further advice and help was provided by James Morison and Mike Dennett, from the University of Reading.

Instat was first used outside Reading on a World Meteorological Organisation (WMO) course in CIMH (Caribbean Institute for Meteorology and Hydrology) Barbados in October 1988. The Climatic Version of Instat and the Climatic Guide were then produced in 1989 and Instat became used by Meteorological Services in many countries, particularly for agroclimatic analyses.

The Climatic Guide was translated into French and Spanish in the early 1990sby staff from the Algerian and Costa Rican Met Services. The climatic facilities of Instat have been used on courses at Reading and on regional courses organised by WMO, Geneva in, for example Burundi, Panama, Philippines, Syria, Algeria, Niger and Brazil. Since 2000 they have been used on regional courses in Kenya and Niger. The software has also been used routinely by the climatological sections in many Met services for the analyses of their historical climatic data

We also acknowledge support from the UK Met. Office over many years. This included the resources to upgrade the climatic facilities in Instat to Windows.

Since 2012 staff at the University of Reading has developed the PICSA approach (Participatory Integrated Climate Services for Agriculture) . This makes extensive use of the historical climatic data and Instat has been used routinely by National Met Services for this work. It will be superceded by R-Instat from early 2017.

FAQ

Frequently asked questions

Introduction

In this section, we tried to imagine ourselves in user's situation. We posed questions we were asked and others we feel that you may have, as you considered whether to install and use the previous versions of Instat. Many continue to be relevance!

General questions:

- What is R-Instat?
- Why should I use R-Instat?
- What about documentation and help?
- So, where is the catch?
- Is R-Instat only for small data sets?

We currently use a spreadsheet (Excel, Open Office) for our statistical work...:

Other Statistical Packages

Finally

FAQ with answers

What was Instat and what is R-Instat?

The original Instat is a general statistics package. It is simple enough to be useful in the teaching of statistics and has the power to assist research in any discipline that requires the statistical analysis of data.

Instat was written by experienced statisticians engaged in teaching, research and consultancy, in collaboration with professional programmers.

Instat is used by many people who have not used a statistics package before. It is easy to learn and to use. It is designed to encourage good statistical practice and to support the teaching of statistics, both in class-room situations and on an individual basis. We hope this will also apply to R-Instat

The Windows version has been developed from Instat around 2000). This was an improvement on our first shareware version released in Autumn 1994. Before then, Instat had been sold commercially. R-Instat is a total rewrite, and is build on the enourmous power of the R statistical system

---

Why should I use R-Instat?

Some points are as follows:

1. It is simple to use.
2. All the documentation is available on-line.
3. It encourages good statistical practice. For examples of what we mean by this, see the series of Case Studies or the Introductory Guide.
4. It is quick and easy to install. Installation usually takes less than 5 minutes.
5. It has a wide range of facilities. These include:

  - good data manipulation
  - descriptive statistics
  - plotting
  - multi-way tables
  - simple, multiple regression, gelenralised linear models, etc
  - analysis of variance
  -
- It can handle quite large data sets, for example
  - 10 variables by 200 cases is no problem
  - 100 variables by 10,000 cases is now easy
  - We are keen to see the limits of R-Instat
- Instat was free for individual non-commercial use (now it is completely free). R-Instat is both free and open source. It can be downloaded and includes both the program and the full documentation. It can easily be distributed to friends, colleagues, or students.

What about documentation and help?

One problem with the effective use of packages can be the lack of documentation when it is needed. This is particularly the case with shareware packages. This is not the case with Instat. There are 4 main guides for Instat, as follows

Tutorial, Introductory, Reference, Climatic

This version includes all these guides. They are available on-line. When you install Instat you can automatically install all these guides. This means that if you are teaching a group of 20 students, all of them have easy access to the manuals. The guides are available as help files and in printable pdf format.

Both the program and the Guides can be downloaded from our web site http://www.reading.ac.uk/ssc

So, what is the catch?

There isn't a catch. Instat has been written and used by staff who are interested in improving the teaching of statistics. Statistics is often not a popular subject. This is sometimes because courses are unnecessarily theoretical, particularly when taught to students for whom it is not their main subject.

We are often amazed to see a statistics course still taught separately from computers. In some developing countries, computers are still not plentiful for individual students, but where they are available, their use can make a statistics course much more relevant and more fun.

In the help files on Teaching and in the new Introductory Guide, we describe ways that statistics teaching can develop, once computers and statistical software are available. We also include a set of 11 Case Studies, as well as many examples from R packages, etc.

We believe R-Instat can continue in its role of providing support for the teaching of statistics, and for research that involves statistical analyses, for many years to come. We hope you enjoy using it.

Is R-Instat only for small data sets?

No.

It also depends what you mean by small. Think of your data as one, or more rectangular worksheets, or data frames. In each worksheet the columns are the variables you measured and the rows are the cases. In the old Instat the maximum number of columns in a worksheet is 127. Now, the front-end grid apparently has a limit of 32,000 columns - though the limit in R is higher!

For the maximum number of rows, i.e. cases, lets take some examples. With experimental data there may be 100 rows, and that is low, and so fine. With daily climatic analyses for 50 years there are close to 20,000 rows and that is fine too.

If you have 50 stations, each with 50 years of daily data, then you are close to 1 million rows. That is the limit in many spreadsheets, including the grid we use as the front end to R. However, the grid is just a visual window to the R data frames, where all the work is conducted. Well before having 1 million rows we suggest you don't have all the data displayed in the front-end grid. R itself can cope with much longer data sets.

So, the limits are very high. However, processing the data, with very long columns may take a long time. We suggest the limits now are concerned with processing time and not space. If you have a very large data set - with millions of rows - then we strongly suggest you migrate from using R-Instat to using R itself. And if you really want to know about large data sets, then R has a task view that can help. It is here: https://cran.r-project.org/web/views/HighPerformanceComputing.html

We currently use a spreadsheet (Excel) for our statistical work, so should we move to Instat?

No.

Using any statistics package does not mean abandoning Excel. Consider instead adding a statistics package to your use of Excel.

Most statistics packages including Instat can read from Excel workbooks and also write to Excel. So you can still continue doing data manipulation, tabulation and graphics in Excel. If you add a statistics package, then you could just use it for the more advanced statistical work that is not in Excel or is poor in Excel.

Where does Instat fit in relation to a spreadsheet?

We are enthusiasts for Excel, even for statistical work. Excel is useful, because of its powerful data manipulation, tabulation (pivot-tables), and graphics. Its dynamic nature (the way it updates automatically) and ease with which you can add to Excel are excellent.

However there are two problems with Excel for statistical work. The first is that Excel does not encourage "good statistical practice". We think that this can be put right, but it does need an add-in. See our web site for further information.

Second, it has weaknesses once you consider more advanced statistics than simple tables and graphs. So, when you fall off the end of Excel for statistical work, you need to add something. This could be one of the many add-ins. It could be a (free) statistical package like Instat, (which can be downloaded from www.ssc.rdg.ac.uk/), or it could be a commercial statistics package.

Instat does encourage good statistical practice. It supports the teaching of statistics to a reasonable level. It may be sufficient for all the analyses you need. If it is not sufficient, it can be a useful "stepping-stone" to encourage you to add one of the commercial statistics packages to your repertoire.

Couldn't Instat be an add-in to Excel?

We do have an add-in to Excel, to encourage "good statistical practice" when using Excel. It would have been too difficult to make the current Instat into an add-in.

Maybe if it is very successful, then a future version might be usable as an add-in? For now, Instat's importing from Excel is very flexible and it can write data to an Excel workbook.

How does Instat compare with Excel add-ins?

Add-ins range from free, to 10's of pounds (dollars) to some at many hundreds of pounds. They are not to be ignored, because they can add useful features, like boxplots or improved regression. We think that Instat compares well, with all but the most expensive. And if you are looking to spend 100's of pounds, then we suggest that you are likely to opt for one of the standard commercial statistics packages. They have also become easy to use.

How does Instat compare with other free software for statistics?

We are obviously hopelessly biased, and the situation keeps changing. So you should look for yourself. A web site with comprehensive information on statistics packages is http://www.ltsn.gla.ac.uk/ from the old Computers in Teaching (CTI) initiative.

How does Instat compare with commercial statistics packages?

The basis for our comparison is that we use many standard packages in our statistics courses. For our own students these include Excel, Minitab, SAS, Genstat and S-Plus. In our Statistical Services Centre short-course programme and commissioned courses we have also used SPSS, JMP, Statgraphics, Statistica, StatExact, Conoco or MLWin.

On the statistical side, with the exception of Excel, Instat is not as powerful as these packages.
Instat is intended to complement, rather than compete with, the standard statistics packages. For those who currently just use a spreadsheet, it can be a "stepping-stone". Once you are familiar with any Windows based statistics package, you will find it easy to add another.

If you are already using a standard statistics package then consider adding Instat, because of its use as a teaching tool. Most standard packages are primarily to process data, though they can be used to support the teaching of statistics. Instat is the other way round. It is largely to support the teaching of statistics, though it can be used for data processing.

What do you mean "Not as Powerful?"

Modern statistical methods have advanced particularly in two areas. The first is the processing of non-normal data and the second is the processing of data at "multiple levels".

For example for analysing data from "contingency tables Instat includes facilities for log-linear models, that is much more powerful than the traditional simple chi-square test. That is an example of the analysis of data from a non-normal distribution. We even include a guide to explain the comparison in detail, because this sort of development is to be encouraged. The standard statistics packages have log-linear models and much more.
If you have data at different levels (like people within households) you might need software that can analyse these data at the multiple levels. If you have such data, then look first at our analysis guide (see our Help menu), because you don't always need to do the analysis at the different levels. If you do, then you need, not just any commercial package, but a powerful one, like Genstat (with REML), SAS (PROC Mixed), S-Plus or MLWin.

What is your relationship with commercial statistical software producers?

In summary it is "good, but not commercial." We do not act as an agent for any producer, nor to we receive any commission on sales.

We can therefore give unbiased advice on statistical software. As we give Instat away to all individuals, we feel that it complements, rather than competes with the standard statistics packages.

Why does Instat have a teaching menu?

In this version we have moved the Teaching Menu to be part of the Help system.
We describe what this means in the Teaching help file.

Our main aim is to encourage good teaching practice, rather than Instat. We believe imaginative teaching requires the use of a software package, but it doesn't have to be Instat. It also needs other resources. Those from Reading are freely available, whether, or not, you use Instat. We also list web sites where you can find resources elsewhere.

Why does Instat have a climatic menu?

This has been a research area at Reading for many years, and Instat has included facilities for the analysis of climatic data for 15 years. We believe Instat is currently the only general statistics package to add these facilities. They were a main reason that the DOS version of Instat continued to be used and led to support from the UK Met Office for the development of the Windows version. We are very grateful to the UK Met Office for this support.

What are you, as producers, hoping to get from Instat?

"Hope" is the key word, and we are hoping for a lot. We usually do, until realism sets in. Our hopes include:

- More imaginative use of the computer to support the teaching of statistics.
- Improved teaching of statistics (and analysis of data) generally, and particularly in developing countries.
- Publicity for the skills of the SSC in training, software evaluation and software development.
- Funds from department and site licences to support further developments of Instat.
- More constructive and effective analyses of climatic data.

Getting Started with R-Instat

System requirements

Still to add - including types of machine it will run on.

Facilities in R-Instat

To be updated for R-Instat

Instat+ is a simple general statistics package. You can enter and manipulate data, summarise data, plot graphs and perform t-tests, regression analyses and analysis of variance, etc. For more details see the Instat+ Introductory Guide
.
Instat+ also has a range of special facilities for the analysis of climatic data, which are described in the Climatic Guide. These include the summary of daily data into a weekly, 10-day or monthly basis, calculations of the dates of the start and end of the rains for a wide variety of definitions, water balance, Penman, degree days and so on.

There is a special Teaching sub-menu, where we describe ways in which access to computers, together with statistics packages like Instat+, can improve both the way standard statistical is taught and also how courses can be broadened. For example, most standard courses do not show how to handle large data sets. The user is then poorly equipped to deal with many of the real problems of statistics.

The Introductory Guide includes ideas on the teaching of statistics and gives suggestions on alternative strategies for data processing.

We try to encourage good statistical practice, and also include a series of good practice guides, within the teaching menu.

If that is not enough, look at frequently asked questions and history of Instat.

R-Instat Windows

The two main R-Instat windows are as shown below.

Add captured picture

The Output window, shown above, displays all the R-Instat commands and their (non-graphical) results. The output is generated from R-Instat's dialogues.

A Current Worksheet window is also shown above. This provides a spreadsheet view of the set of data in the worksheet that is being analysed.

Other available windows are:

A Log window that keeps a log of the analyses you have run in the current session.
A Script window that is used to record specific R commands.
One or more Graphics windows to display the graphical results.

Help is displayed in its own Window and is available either from individual dialogues or from the Help menu.

Using dialogues

R-Instat, in common with other Windows software, is used mainly through a system of menus and dialogues.
We first look briefly at the role of each part of the menu system and then describe the common features of the dialogues.

R-Instat's top menu is as follows:

Add captured picture

The menus File and Edit on the left and View and Help on the right, are common to most software.

All statistics packages have menus that reflect the need for data management or manipulation - R-Instat has Organise, Describe and Model menus.

That leaves the Climatic menu, because R-Instat is also a package to support the analysis of climatic data.

We describe each of these menus in turn and then each dialogue. This is our top-down approach! You may alternatively start with the help on an individual dialogue and climb the tree instead.

First the common features of R-Instat dialogues.

To access most of the R-Instat dialogues you first open an Instat object or import a file of data. They are stored in data frames in R. Each data frame is like a sheet in a spreadsheet package, and sometimes we will call them sheets.

You can have many sheets open at the same time. Then, like Excel, the name of each sheet is given in a tab at the bottom.

When you first use a dialogue, the OK button is usually not enabled. You must first supply the minimum information, usually the column(s) of data to process.

In dialogues where results are to be stored, in the current or another sheet, then R-Instat suggests a name for the new columns. If more than 1 then it suggests a prefix. You are at liberty to change these names.

Although R-Instat will never suggest, you are welcome to overwrite the information in existing columns.

Once you complete a dialogue and press OK, R-Instat does a few checks and generates the appropriate R command, that it sends to the R for processing. While the server is working you will see Busy on the status line at the bottom of the R-Instat window.

The results are then usually returned to the Output window. If there was an error in the command, you will be informed, both in the Output window and through a message.

Getting help

There are many ways to get help. We assume that users will often "browse" through the dialogues and look for Help from there. Each R-Instat dialogue has a Help button. This gives direct information on the dialogue and also will usually contains 3 special keys, labelled

Commands Examples Data

as can be seen in the help for the Manage => Calculations dialogue help.

Commands lists the R commands that correspond to the dialogue and you can jump from there to the appropriate place in the Help.

Examples will describe where the dialogue is used, usually in the Introductory, or Climatic Guides or in the Case Studies. You can transfer there if needed for more information.

Data will tells you which of R-Instat's sample datasets can be used to illustrate the working of the dialogue. You can jump to a fuller description if needed.

The main alternative to the dialogues is the Help menu, which gives access to general information and to the contents (more than 1000 pages) of all R-Instat's documentation. Here are 3 ways to use this:

1. Choose Help => Contents.

2. For help on how to use R-Instat, select Help => Getting started in R-Instat.

3. To see, for example, how to analyse your data using ANOVA, you will be able to choose Help => Introductory Guide and go to Chapter 16.

Ways of using R-Instat

To be updated from Instat

Here we will describe alternative ways of using R-Instat. To give instructions to R-Instat, you can use dialogues, commands or a mixture of the two.
We first describe the different options and then consider different stages, such as data entry, manipulation and the presentation of results. If you wish to try these alternative methods, open a data file, File => Open Worksheet and select Survey.wor.
1: Using dialogues
If you wish to display a histogram for the data in X6, use Describe⇒ Specific ⇒ Histogram, complete the dialogue and click OK. An example is given below.

If you think that this is the only way to use R-Instat, then read "Why not just use dialogues".

To be updated for R-Instat
2: Giving commands in the "Commands and Output" Window
Go to the Session window and type
HISt X6 <Enter>
This window is interactive, so <Enter> is like the OK button and gives the same histogram as 1).

3: Giving commands in an Input window
Use Edit => Edit Text => New or Edit Macro => New to open a new Editing window and type
HISt X6 <Enter>

To run the command, use
Submit => Current Window.
The Input window is not normally interactive, so you can type several commands, each on a new line, and then run them all at one go.
4: Giving commands in an interactive Input window
Use Edit => Edit Text => New as in Method 3 and Submit => Interactive, to make the Editing window interactive (like the Commands and Output window). Type
HISt X6 <Enter> to produce the histogram.
Why not just use dialogues?
You may wonder why anyone would consider anything other than using the dialogues. They certainly do away with the need to know all the syntax of the R-Instat commands. In fact, what happens with the dialogue is that when you press OK the command "HIST X6" is constructed for you and sent to the R-Instat server. So, to the server, all the methods are the same.
If you are a beginner or use R-Instat infrequently, we suggest you always use dialogues. One possible difficulty with dialogues is getting help from other people if things go wrong. However, a record of the commands you 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 second 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 the mouse clicking is time consuming. It is also very odd, because computers are particularly good at repetitive tasks, so why are we doing all this repetition!
You could alternatively just use the mouse the first time, keep a record of the commands that were generated in the Log window, and submit them to automate the analysis for the other 19 data sets.
Comparing different ways of giving commands
Typing commands into the Commands and Output window will be familiar to anyone who used a statistics package in DOS. There was usually just one text screen and this was used for everything.
We suggest that typing the commands into an (interactive) input window is just as simple to do and will be clearer when you wish to review the analyses you have done.
However, if you wish to just use the Commands and Output window, we suggest you also keep a separate record of your commands in a Log window.
Once you are in a routine of using a separate Input window, it is a small step to collect a few commands together before submitting them all to R-Instat in one go. This is probably useful if you will have further sets of data to analyse on another occasion.

See Macros in R-Instat to find information on writing files.

A typical analysis

A typical R-Instat analysis includes the following stages:
Data entry or retrieval
Data management
Data exploration
Graphics
Data analysis
Saving and printing your work
Getting Started - Data entry or retrieval
Choose one of the following, from the File menu:
File => New Worksheet - to enter a new set of data
File => Open Worksheet - if your data are in an existing R-Instat worksheet
File => Open Worksheet - if your data are to be imported.
Any of the above methods results in a worksheet being opened in the R-Instat Windows server. The worksheet is displayed in a similar format to a spreadsheet in, but with certain limitations, in the Data window.
To enter new data, type it into the appropriate cells of the worksheet.
R-Instat is also supplied with many sample data sets. These are usually in the folder called Intatws and are called *.WOR. Four files that are used in the R-Instat Introductory Guide are
SURVEY.WOR data from a simple survey with 36 rows and 6 variables
EXPERI.WOR data from a simple experiment
REGRESS.WOR data to illustrate simple regression
RAIN.WOR 32 years of monthly rainfall records
Getting Started - Data exploration
In processing data, it is convenient to distinguish between data exploration and data analysis, although these two stages are usually conducted together.
Data exploration involves looking closely at the data to explore both the pattern in the data and any possible oddities. Graphs are an important part of this process and R-Instat facilities for graphics are described in the Introductory Guide, Chapter 11.
Data can also be presented in tables as described in the Introductory Guide, Chapter 13.
Getting Started - Data management
In simple "text book" statistics, the data that are entered are already in a form that is needed for a particular analysis.
This is rarely the case with real data and is one reason that users often find that "real data" are confusing to analyse.
Thus real data often have to be manipulated, so that they are in a suitable shape for the analysis. This step includes transforming the data, so that they are in the correct units. It may include re-coding and sorting some of the columns or even changing the shape of the data.
Most of the facilities in R-Instat for this management step are in the various choices from the Manage menu. Additional facilities are in the Statistics menu, if an initial summary is needed, prior to the analysis.
Getting Started - Graphics
The range of graphics is shown in Introductory Guide, Chapter 11. The difference is that the Graphics => Plot and Graphics => Histogram, dialogues use the Visual Basic front end, to do the plotting. This means that more colours are available and you may have multiple plots in separate Windows at the same time. The Window => Graphics menu gives various options on tiling or cascading the graphics Windows.
Getting Started - Data analysis
R-Instat provides a wide variety of ways in which your data can be analysed. They are all available through the options of the Statistics menu.
Getting Started - Saving and printing your work
Results from the output window can be copied to the clipboard and pasted into reporting software, such as Word, as you would expect from Windows software. You may also switch a secondary output window called a spool window (Edit => Output Spooling) to keep a copy of just the important results that are needed in a report.
The current graph can also be copied to the clipboard, using Edit => Copy or <Ctrl+C>. It is copied in its entirety, you can not select part of the graph. If you wish to do this, then you will have to use special

software, such as PaintshopPro. Note that the copy will be of the graph as viewed and the results may be clearer if you maximise the graphics window before copying.
Once in the clipboard the file can be pasted into Word, or other software, using Paste Special and choosing to copy the graph.

Climatic analysis

For the past 30 years, Instat has been distributed to the Met. Services in many developing countries by the World Meteorological Office (WMO). Instat has been the basis of a series of regional courses on Agricultural Meteorology, organised by WMO. It has also been the main package on the 10 week course called "Statistics in Agricultural Climatology", which has been run since 1986 at The University of Reading, with support from the U.K. Met. Office.
The Climatic menu indicates some of the climatic facilities in R-Instat.

The Climatic Guide is part of the Help system and is available as a pdf file. An indication of the wide range of analyses that are possible for climatic data is shown by the titles of the different chapters of the Climatic Guide.
Chapter 1: Introduction
Chapter 2: Practical Introduction to INSTAT
Chapter 3: Setting up a Worksheet and Entering the Data
Chapter 4: The Presentation of Climatic Data
Chapter 5: The Summary of Climatic Data
Chapter 6: The Start of the Rains
Chapter 7: Dry Spells and Water Balance
Chapter 8: Analysis of Temperature Data
Chapter 9: Agricultural Climatology
Chapter 10: A Crop Performance Index
Chapter 11: Further Topics
Chapter 12: Case Studies
Chapter 13: Modelling Daily Rainfall Data
Chapter 14: Writing Climatic Macros
Chapter 15: Conclusions

The main role of R-Instat is to complement the standard analyses that are often undertaken by Meteorological services. R-Instat encourages a flexible approach to the analyses, where the characteristics to be investigated are dictated by the user, given the objective for which the analysis is being done.

The main purpose in producing R-Instat has been to continue to support users in Met. Services to exploit their climatic data to the full. Users now demand an ease of use that is possible by exploiting the facilities of a Windows environment.

Data processing strategy

We assume that you will be using R-Instat for the analysis of your data, though the ideas in this section apply whatever statistics package you have. Data processing is more than just analysis. You also have to decide how you will:

- Enter or transfer your data
- Manipulate the data
- Present the results

For example, do you propose to enter your data directly into R-Instat, into a spreadsheet or into a special data-entry system? If you are happy manipulating data in a spreadsheet, do you propose to continue, or start to do the manipulation in R-Instat instead? In this section we give some guidance on such issues.

Data entry

Data entry

We assume first that the data are on paper. See the section on data transfer if the data are already computerised.
To define your method for data entry, ask yourself the following questions.

Do you have much data?

6 columns and 20 rows is not much;
50 columns and 3,000 rows is quite a lot;
250 columns and 80,0000 rows is a lot - don't use Instat. Many other statistics packages can cope, and even with much more than this. Check what to say now?

Are your data simple or complicated?

Simple data are examples that can be visualised easily in a single rectangle, e.g. 6 columns and 20 rows.

Complicated examples are when data are available from more than one 'level'. Examples at more than one level are a survey with data at village, household and person level, or an experiment with yields recorded at plot level and insect counts at plant level. Complicated data sets are best visualised as a set of (linked) rectangles, with one rectangle for the data at each 'level'.

One possibility is to enter your data directly into the data frame (spreadsheet) in Instat. Consider this option only for small and simple data sets. Sometimes some of the columns are a regular sequence, like 1 1 1 2 2 2 3 3 3 4 4 4. R-Instat has neat ways of entering such columns.

What the current version of Instat does not yet have is built-in facilities for data verification. So, an alternative strategy is to consider a package that is better designed for data entry.

A popular choice is a spreadsheet package such as Excel. This is possible, but be careful. Excel can be used effectively, for data entry but it needs to be used with 'discipline'. If you are considering using a spreadsheet package and are not sure what 'discipline' means for your work, then look at the 'good-practice booklet' entitled 'Disciplined use of spreadsheets for data entry', which is also on the Statistical Services Centre web site (http://www.reading.ac.uk/ssc). We also have training course notes on the effective use of Excel for the entry of (experimental) data in both English and French if these are required. See our Web site again.

If your data set is complicated (i.e. not at a single level) then consider a database package, such as Access, for data entry and management, rather than a spreadsheet, particularly if the data set is not small. We have a good-practice guide again.

EpiInfo and CSPRO are free and useful for data entry, particularly for survey data. They are options for complicated data sets.

A few statistics packages do add facilities for data entry. SPSS has a data entry module, which is a separate product, while Genstat includes facilities for double entry of data.

Data transfer

In this section we assume that your data have been entered into a different package and need to be transferred to R-Instat for analysis.

If they are in the correct shape, then the transfer should be easy. Use File => Open to import into an R-Instat data frame and start analysis immediately. Importing can be from Excel, or from another statistics package etc.

There are various ways that data can be awkward. It may be the wrong shape, perhaps columns and rows need to be interchanged or it needs to be 'stretched' out. (If you are not sure what 'shape' data sets should have for a statistics package, then look at the examples described in the Introductory Guide, Chapter 3.)

The data may be interspersed with results. One common fault in entering data into Excel is to confuse data entry with the presentation of summary values (like means). In this case the data are also often the wrong shape as well!

The data may be 'fixed format', so there are no spaces between some of the numbers.

If your data are 'awkward' then do not expect a 'free ride'. You will have to think carefully on the steps to be taken to import the data into IR-nstat or any other statistics package. With 'awkward' data the first question is 'Do you have just one set of data or are there many?'

If it is just a single data set, then sorting it out interactively with a package such as Excel is a possible option. If there are many data sets, then it is preferable to devise a routine that can be applied automatically, even if the importing takes a little longer for the first set. If you use Excel, this might mean writing a macro.

Some guidelines to help you devise an appropriate strategy are as follows:

1. Allow enough of your time to import the data sensibly

2. Consider whether you can separate raw data from summaries. If they are in a spreadsheet, perhaps part of the sheet with the data can be a 'named range'. Perhaps all the raw data can be copied to a different sheet in the same workbook.

3. A common fault is to enter data into many small rectangles, rather than in what Excel calls a 'list format'. An example is in Section 3.8 in the Introductory Guide. One way to change the shape is to import the data rectangle, by rectangle, (ignore the margins) and then use the ***Manage*** ⇒ ***Reshape*** ⇒ ***Stack*** facility.

R-Instat can read data from an ASCII file that is in a fixed format (see the File => Import/Export => Import ASCII dialogue.) Alternatively you may wish to import into Excel, which has good interactive facilities for reading fixed format files.

Decide which package to use for data manipulation, because this may be needed as part of the process. If you are comfortable with R-Instat, then import first and manipulate the data within R-Instat, otherwise manipulate the data into the correct shape with Excel etc.

Data manipulation

In simple 'text book' statistics, the data that are entered are already in a form that is needed for a particular analysis.

This is rarely the case with real data. Then, users typically adopt one of three strategies:

1. Summarise the raw data using a calculator, so that the data are input into the computer ready for analysis. Avoid this - see the Data Management Guide if this is your current strategy.

2. Enter the rows of data into a spreadsheet package, such as Excel, or a database package, such as Access or a special data entry system, such as EpiInfo. Then manipulate the data in this package and export the data to a statistics package in the correct form for the analysis.

3. Enter the raw data directly into a statistics package. Then manipulate the data, within the statistics package, so they are in a suitable shape for the analysis. This step can include transforming the data, so that they are in the correct units. It may involve re-coding and sorting some of the columns or even changing the shape of the data.

As with data entry a key distinction is whether the data are simple or complicated, i.e. are they at one level or at more than one level. If at a single level, i.e. a single rectangle of data, then Excel, or R-Instat is fine for the data manipulation. If they have a more complicated structure, then consider software, such as Access, or Epi-Info for the manipulation stage.

R-Instat's manage menu shows the facilities for data manipulation in a typical statistics package. Most are for data at a single level, but the Manage => Reshape => Expand shows how data can be moved down from one level to another. The converse is to summaries data at a lower level and then use the summaries in a further analysis. This uses the Statistics => Summary => Column Statistics or Statistics => Tables dialogues and is the equivalent to Pivot Tables in spreadsheet software, such as Excel. Rewrite, because this version of Instat may be OK.

Analysis

Instat's facilities for analysis are mainly in the statistics menu. They are comprehensive and will provide all the facilities needed for many users, particularly those who have been trying to do their analyses using a spreadsheet.

One role for Instat is to show what is simple in a statistics package, that is awkward and time consuming with software, such as a spreadsheet, that was not particularly designed for data analysis.

A guide on analysis describes our strategy for the analysis of experimental data, and there are other guides on the analysis of survey data and general modelling.

A key concept in data analysis is that all real datasets have structure and the analysis should reflect this structure. The structure may be simple, for examples different groups of data (we use factor columns to indicate the groups). More complicated structure involves data at different levels, as mentioned above.

Analyses sometimes suffer, because researchers have not fully grasped the concepts of statistical inference and again we have a good-practice guide. Instat is particularly strong in supporting the teaching of these concepts, see, the Statistics => Simple Models menu.

In planning their strategy for analysis, Instat users need to know the limitations of R-Instat, so they know whether they are working efficiently with Instat, or should migrate to a more powerful statistics package. In any references to other software, we try to be unbiased.

Instat is fine for simple statistics, and is also good for tabulation. An alternative for the latter is to use Excel's Pivot tables, which are very flexible. SPSS has a special tables module. See the Introductory Guide, Chapter 13 for more information on tabulation.

From version 3.1x, Instat's graphics facilities have been improved, both for exploration and for presentation. Details are in the Introductory Guide, Chapter 11.

Instat has quite good simple and multiple regression facilities. If what Instat has is not enough, then make sure that the more powerful package you choose has good facilities for the inclusion of factors in regression models. Some packages publicise facilities for non-linear regression, again make sure they can cope with factors in the non-linear modelling. More information is in the Introductory Guide, Chapter 17.

Instat's facilities for log-linear models indicate the advances there have been in handling non-normal data. This is called generalised linear modelling (more general than general linear models!) and log linear models are just one example. So if you have non-normal data and don't want to transform and hope that all has magically become ok, then look for the statistics packages that can handle generalised linear models. See the Introductory Guide, Chapter 17.11 for more information.

Instat has good facilities for the analysis of experimental data (ANOVA). We are often disappointed in these facilities in the standard packages, sometimes in what they do, and sometimes in the ways they present the results. There seems to be much less competition in this area and only one package stands out. If you need more than Instat, then look at Genstat. The Introductory Guide, Chapter 16 has more information.

A recent major advance has been in the facilities for multi-level modelling. A simple example of an analysis at two levels is a split-plot experiment. Even if you have data at multiple levels, our analysis booklet explains why you may not need to analyse the data as a multi-level model. If you do need to fit a mixed model, then you need more powerful software than Instat.

The SSC uses many packages for training courses and consultancy work. In addition to Instat and SSC-Stat, these include Access, Canoco, Excel, GenStat, JMP, Minitab, R, SAS, Stata, StatGraphics, Statistica, S-PLUS, and SPSS. The SSC does not act as an agent for any commercial software, nor does it receive any financial advantage from the sales of any package.

Presentation of results

Results, following an analysis, are displayed in reports, either as tables, as graphs, or within the text. Our guide called Informative Presentation of Tables, Graphs and Statistics, gives suggestions on what we consider to be good practice in these areas.

It is important that tables of results are copied into a report, rather than being retyped. With Instat, tables can be copied from the Output, or the Spool Window, though they may have to be reformatted in the report.

And more, please!

Discuss archiving and "Completing the circle."

R-Instat Spreadsheet

We expect most users will already be familiar with a spreadsheet. R-Instat has its own spreadsheet "system" and we have included a range of features that we hope spreadsheet "addicts" will find familiar.

We have divided our explanations into three topics as follows:

Similar but Different

Data In and Out

Right-Click Menus

Similar but Different

Data In and Out

Right-Click Menus

Menus and Dialogues

Instat's Main Menu and Toolbar

Detailed information about the menu is under the following headings:

File	Operations on files, such as opening, saving and closing.
Edit	Usual cut/copy/paste operations, plus other utilities.
Organise	Data entry, manipulation and removal.
Describe	Summary statistics, tables and graphs.
Model	Statistical modelling, including regression and ANOVA.
Climatic	Optional menu on the summary and modelling of climatic data.
Tools	Various utilities and options
View	Arrangement of the windows in R-Instat and views of the data frams
Help	Access to the Instat Guides and resources and ideas on the use of statistics to support statistics teaching and learning.

From the menus you choose dialogues which give instructions to R-Instat. Each dialogue has its own HELP and you can step through this information, either from here or starting from any dialogue help.

File Menu

The file menu contains what you would expect. You can start with an empty file (spreadsheet), ***File > New*** or import one that already exists ***File > Open***. We are using the rio package for File Open, and so (in principle) can open files in many different formats.

New Data Frame

Creates a new worksheet for R-Instat, which works differently from most Windows software in that the empty file is created at the beginning in a file in your working folder. Then, the contents of the file are updated automatically, as it is used.

For a worksheet there is therefore no need to save the file when you have finished. This also means that if there is a problem, for example a power cut, while you are working, the worksheet will contain all your work except the change that you are conducting at that moment.

At the start you have to state the dimensions of the worksheet that you wish to create. If the defaults are not suitable, then you can specify the size. Alternatively you can dictate that the dimensions be the same as an existing worksheet. If later you find that the dimensions need to be changed, use Manage => Resize Worksheet.

R command:

Open from File

Opens an Instat worksheet for further analyses. Also imports into an Instat worksheet from a variety of other formats.
You may only open a single worksheet at a time. Opening a worksheet automatically closes the worksheet that was previously open.

Once open, any change in the worksheet is automatically saved to the disc, so there is no need to save periodically. This does mean that you can not undo major mistakes that have already been saved to the disc. The backup option (from Edit => Options) or the use of the worksheet editor are available if this makes you nervous!

Instat also has a facility to Lock columns of data, so they can not accidentally be changed during an analysis.
By changing the type of file, data can be imported from a wide variety of spreadsheet packages and from many statistics packages. In this case the data are first copied into an Instat worksheet, using a special program called Dataload, and then this Instat worksheet is opened. When importing there will be a second dialogue that asks for the name that you wish to give to the Instat worksheet. This importing is done only once. On future occasions with this file, you would open the Instat worksheet that was made.

When the transfer is from a spreadsheet, the default is to take the whole of the first sheet. From Excel, there are options to choose otherwise. More information can be found in the Import dialogue

---

R command:

Open from Library

This is the same as File => Open, except that the folder is set to the library of datasets supplied with Instat. The file from the library may either be an Instat worksheet, or may be imported from a different file type.

When the library file is an Instat worksheet, then a copy of the file, with the same name, is made in your current working folder. In future use this copy, by using File => Open Worksheet. You only use the library copy, the first time that you need the worksheet.

When the file is imported from a different format, you will be asked what name you wish to give to the Instat worksheet. Normally keep the same name. Once imported, use the File => Open Worksheet on future occasions, to continue with the Instat worksheet that was made.

The library copy thus remains as a "backup". You only need to return to the library for the same file, if you need to return to the version as supplied.

R command:

Import from ODK

Import from CSPRO

Import from Databases

Convert

Save

Save As

Data

Output

Log

Script

Export Data

No dialogue for this yet and it all may be handled from File Open.

The Import ASCII allows text data to be imported into an Instat worksheet. Similarly the ODBC Query allows the data resulting from a query to be imported.

The formats for Export as… are more limited, but they do include export to Excel, and most other software can read Excel files. The Output dialogue is limited to the production of ASCII files, but they can be for a defined subset of the data in a worksheet, and can be in a format that is ready for some of the standard statistics packages.
Alternatively data can be copied and pasted into (or from) an Instat worksheet via the clipboard.

ODBC (Open DataBase Connectivity) is a standard developed by Microsoft to allow a common method of accessing a variety of databases. In Instat the data resulting from an SQL query are transferred into a new Instat worksheet. The ODBC facility is a separate programme adapted for Instat by David Baird and includes its own Help system.

Outputs specified columns of data from Instat to an ASCII file. If the appropriate option is chosen then the output file includes commands so it is easily read into some common statistics packages, particularly SAS (www.sas.com), SPSS (www.spss.com) and Genstat (www.nag.co.uk).

We will probably use Foreign for this.

Print

Print preview

Exit

Edit Menu

Find

Could this include Find Next and Find All?

I assume no replace

Find Next

Replace

Cut

Copy

Copy Special

Paste

Prepare Menu

Data Frame

These dialogues are all concerned with setting up and displaying data in an R-Instat data frame. They can be used for data entry, for generating random samples, displaying data in the Output window and for naming columns.
The most used dialogue is likely to be to enter regular sequences.

View Data

User to display a subset of the data in the output/results window.

Sometimes you may wish to display some of the data in the Output window.

R commands:
head
tail

Rename Column

Usually used to enter or rename a single column in the worksheet.

It is also possible to enter a sequence of names, when the first name finishes with a number, e.g. Week4. Then the subsequent columns will continue the sequence, e.g. Week5, Week6... .
Names in R-Instat follow the same rules as names in R. They cannot contain a space (use _ instead) or arithmetic characters, such as -, =, +. And they must start with a letter or a full-stop.

This dialogue is also available by going into the column name in the data frame, using <Right Click> and choosing Rename.

R command:

Row Numbers/Names

Sort

This is a simple dialogue to sort a column of data and carry other columns along with this sorted order. It is a multiple sort, so if there are ties in the first column, then these are sorted in the order of the second, and so on.

You may choose whether to sort in ascending (the default), or in descending order.

R command: order

Filter

Need to add help!

R command:

Replace Values

Convert Columns

R command:

Column Metadata

Need to add help!

R command:

Column Structure

Reorder Columns

Insert Columns/Rows

Delete Columns/Rows

Protect Columns

Hide Columns

Freeze Columns

Colour by Property

Check Data

Need to add help!

Also need to add dialogues and analyses. This could also relate to a data entry and quality control system.

R command:

Jitter

Export to Open Refine

Import from Open Refine

SDC Micro

Column: Calculate

Calculations

This is Instat's calculator. Examples are in the Tutorial, Section 5 and the Introductory Guide, Chapter 8.2. It has a number of different uses:
It can be an ordinary calculator, presenting the results in the output window.
As such you just type the calculation, roughly as you would write it, for example
7*(3+4) + 2^3 + 8/2 and press OK.

If the result is needed for future use, then click on Save result and choose where to save it. In the above example, it would probably be saved into a constant, say K1, as the result is a single value.

It is a scientific calculator, particularly using the maths functions.

It is a "column calculator", so all the calculations can equally easily be specified using columns. For example Ln(X1+0.5) would calculate a column with the data transformed to logs.

It is a statistical calculator, and the Summary option indicates the statistical functions you can use. You can also include probability calculations for common distributions or to give combinations and permutations.

You can also include probability calculations for common distributions or to give combinations and permutations.
The 5 types of function with the calculator are as follows:
Maths, Logical Summary Probability and Dates.

When you use a function, it will appear in a form that can be edited. For example pressing on the Maths and then the cosine function, it will appear as cos(radians) to remind you to include a number, or a column that is in radians. With some functions there are multiple arguments, for example Probability and then tpr gives tpr(number,df). You can move between these fields by using <Alt> or <Alt> .

See the Reference Guide for further information. Also see Manage => Calculations for the common transformations in statistical work.

Instat remembers the last few calculations, so these can be edited when needed. It starts with some examples of possible calculations.

Maths

The dialogue includes tooltips that give an example and some explanation.

These give the standard mathematical functions, with the addition that they normally transform a column of data, rather than a single value. So Int (4.1) gives 4, while X3 = Int (x2 + 0.1) would give the integer part of each value in x2.
See also Manage => Manipulate => Transform for the common transformations in statistical work.

Note that the functions, such as Sine and Cosine, assume that the data are in radians, so use Sine(rad(45)) and so on, if the data are in degrees.

The Reference Guide gives more details, and some explanation is given by the tooltips.

Logical

In these operations, remember that TRUE is 1 and FALSE is 0.

So (3>2) =1 (because the statement is true). Similarly (3<2) = 0.
x2 = 1 + (x1>0) + (x1>5) evaluates to 1 or 2 or 3, depending on the value in X1
(This is an alternative to Manage => Manipulate => Recode).

x4 = 1+ (x1>0) + 2*(x2>0) + 4* (x3 > 0) evaluates to 1 up to 8, depending on the values in x1 to x3.

x5 = (x4 AND 1) makes x5 = 0 or 1 for odd and even values in x4.

Summary

These each give a simple summary statistic for a column.
For example, with x1 = 1, 2, 9, 4, 7 Min(x1) = 1 and Max(x1) = 9.

If you need more than one summary, then see Statistics => Summary => Describe or Column Statistics.

Note they can have other mathematical functions nested within them. Thus mean(log(x1)) can be done and is the same as saving log(x1) into a column and then taking the mean.

Other functions can also be nested, e.g. log(mean(x1)) is fine.

These functions can act on a subset of a column, for example min(x1(3]5) = 4 in the example above.

Strings

Probability

With these probability buttons, the only complication is that the different options need different numbers of arguments.
The first 8 keys give cumulative probabilities and deviates for the common distributions associated with the normal distribution. Thus
Npr(x) = Probability(X<x), where X has an N(0,1) distribution, e.g. npr(-1.64) = 0.05
Similarly
Tpr(x, n), for a t distribution with n degrees of freedom
Cpr(x, n), for a chi-square distribution with n degrees of freedom
Fpr(x, n1, n2), for an F distribution, with n1 degrees of freedom (numerator) and n2 (denominator).

The deviates give the inverse thus Nde(p) = the value of x for which the cumulative probability is p, e.g. nde(0.05) = -1.64
Similarly for Tde(p, n), Cde(t, n) and Fde(t, n1, n2).

You can include a column, instead of just single values. For example, if
X1= 1, 2, 3, … 20, then Tde(0.975, X1) would give the corresponding values of x for the 20 different t-distributions.

The Statistics ⇒ Simple Models ⇒ Probability Distributions can alternatively be used for the same calculations.

The factorial function Fac(n) can be used, as long as n is not too large, 33 is the largest. It calculates the gamma function and so can be used for fractional values.

Alternatively use the log gamma function, given by
Lga(n) = Ln(Fac(n-1)), e.g. Lga(6) = ln(Fac(5)) = ln(5*4*3*2*1) = ln(120) = 4.79.

This can be given for larger values, e.g.Lga(201) = ln(Fac(200)) = 863
Per(r, n) = n!/(n-r)! is the number of permutations while
Com(r, n) = n!/(r!*(n-r)!) is the number of combinations.

With these functions, any of the arguments can be replaced by a column, Xn.

The final 2 functions require a column as the argument. Then Cus(Xn) gives the cumulative sum, while Dif(Xn)give the difference between successive elements in the column.

So, with X1 = (1,2,4,7,11)
Cus(x1) gives (1,3,7,14, 25), while Dif(X1) gives (1,1,2,3,4).

Dates

The function date(day,month,year) gives a numeric value for any date between the year zero AD and the year 9999. The value given is the same as is used by Excel. So the baseline is 31 December 1899, which has the value 1. Dates earlier than this have negative values, e.g. date(1,1,1899) = -363.

A column of dates can be displayed as such, using Manage => Column Properties => Format and choosing one of the date formats.

This date function is used if you already have columns containing the day, month and year. Alternatively the Manage ⇒ Data ⇒ Regular Sequence dialogue may be used to enter a column of dates directly.

Once you have a column of dates, then most of the other date functions extract components of the date. For example Month(date) gives the month number, with January as month 1. Similarly:
Year(date) is the year number
Dom(date) is the day of the month
Dek(date) is the dekad (or decade) number.
These month, dekad, week, pentade functions all assume that the year starts in January. If your analysis needs a different starting month in the year, then use the Climatic => Manage => Make Factor dialogue instead. That dialogue also permits label columns with the month, or other names to be attached automatically.

Further information is in the Reference Guide. The date functions can be nested. For example the dow function gives the day of the week, from 0 to 6, with Sunday as day 0. It needs a date, which could be given as dow(date(15,10,2004)).

Duplicate Column

Transform

This dialogue is partly an alternative to Manage ⇒ Calculations and provides a range of simple ways to transform data prior to analysis. An example is in the Introductory Guide, Chapter 8.2.

The dialogue is in 2 parts. The first part provides the calculations that are commonly needed in statistical work. There are different transformations for non-negative variables, for percentages (or proportions) and for circular data.

Following analysis, the summary statistics are normally back-transformed for presentation. Hence the dialogue has an Inverse check-box.

The second part of this dialogue has 6 options. They are to rank a column, to sort, to give normal scores, to standardize, to lag or to give differences. The last two options are sometimes used in time series analysis and an example is at the end of Section 19.3 in the Introductory Guide.

R command:

Rank

Later if ranks are within groups then the plyr package makes this easy.

R command: [rank]

Polynomials

Generates polynomial terms for a variate, x.
The simplest is just to generate x, x^2, x-cubed. The second possibility is to centre the terms. Then the first becomes $(x-\bar{x})$, then $(x-\bar{x})^2$, and so on.

The third possibility is to generate orthogonal polynomials. Then the successive terms are made to be mutually orthogonal.

With the orthogonal polynomials, if the available data column that is chosen is a factor column, then you may specify a further column containing the levels. For example, if a factor column, with 5 levels would be coded with one of the numbers 1, 2, 3, 4, 5 for all the rows of data. You could have a further X or a label column, L, with the levels of the factor, for example 0, 1, 2, 4, 8. Then the polynomial terms would be calculated for this column and expanded to the length of the factor column.

When using orthogonal contrasts with ANOVA, then make the available data column the small X column containing a single value for each level. See the Introductory Guide, Chapter 16.8 for a description of how to incorporate these terms into an Analysis of Variance.

R command: [poly]

Row Summary

A statistics package is primarily a column calculator, but occasionally summary statistics are required for each row of the data. This dialogue provides for the standard statistics. An example is in Section 8.3 in the Introductory Guide.

You may choose to restrict the data values that are counted in the calculations. For example they could be restricted to all positive values.

You can include a column or set of weights to give a weighted mean, or sum. This is often used to calculate an index, and an example of the calculation of an "assets" index is given in the Introductory Guide Chapter 8.4.

When using weights, the count returns the sum of the weights, (omitting and restricted or missing values). The mean is the weighted total, divided by the sum of the weights.

The Index check-box may be used to standardize the resulting weighted sum. If an observation number is given, then the resulting index is standardized, so that row number has an index of 100. Alternatively you may give a value, and the resulting weighted sum will be divided by that value (and then multiplied by 100). Examples are in the Introductory Guide, Section 8.5.

R command: apply

Column: Generate

The dialogues in this menu add further columns, with the manipulated data, to the chosen data frame.

Regular Sequence

Used to enter a column that is regular, for example:
1, 2, 3, 4, 5, 6
1, 2, 1, 2, 1, 2

It is also used to enter a sequence of dates. With this option set, the pull-down list provides a calendar, from which the start date can be chosen.

When a sequence has been specified, the preview at the right of the dialogue box indicates the sequence that will be generated. Check also that the length of the column, i.e. the number of items, reported in the top right-hand corner of the dialogue, corresponds with the length of the current data frame.

R command:
seq

Count in Factor

Enter Column

Random Samples

Permute Rows

Need to add help!

R command: sample,
set.seed

Column: Factor

Convert To Factor

R command: factor

Recode Numeric

Usually used to recode a column with ordinary data, like ages, into a category column, like "child" "adult" "senior". An example is in the Introductory Guide Section 8.3.

Factor columns do not have to have values from 1 upwards, but often it is convenient to recode data into the consecutive integers.

If a single value is recoded, then the "Range end" field may be left blank.

You may give factor labels as the new values. Otherwise, if you choose to give a labelled factor, Instat will add a default label. For example if 18 to 65 is coded as 2, and a label is added, it would be "18-65".

You may also recode an existing factor. This will often be to merge values from different levels. However this is often simpler through the Manage ⇒ Factor ⇒ Recode dialogue.

R command: cut

Recode Factor

R command: recode

Combine Factors

R command: interaction

Dummy Variables

Generates indicator, (or dummy variables) for a factor column. This is usually for teaching purposes, because factors can be included in regression models directly.

Usually you would give n-1 indicator columns for a factor with n levels. The default is to leave the last level as the baseline. But you may choose to give all n levels, or to dictate which level is to be taken as the baseline.

An alternative is to use the Statistics => Regression => Simple with Groups dialogue, which generates these indicator columns automatically. These may then be kept to show what has been done.

The dialogue also allows for a variate to be included, so that parallel regression lines can be compared with a "full model". If you use this option you would normally use the dialogue twice, first to get the dummy variables and then, with the same factor, to specify the variate of interest.

Indicator variables can be given for more than one factor, enabling Instat's regression to be used more generally. This is done automatically in the Statistics => Regression => General Linear Models dialogue, and illustrated in Section 17.6.
With the availability of the Simple with Groups dialogue and the General Linear Models dialogue, this Indicator dialogue is mainly for teaching purposes, so users can understand how regression models are fitted when factors are involved. It can also be used in a macro.

R command: dummy

Levels/Labels

R command: factor

View Factor Labels

Reorder Levels

Reorders the levels (labels) of a factor column.

R command: factor

Reference Level

R command: relevel

Unused Levels

Also shows the frequencies at each level of a given factor.

R command: droplevels

Contrasts

Factor Data Frame

Column: Text

Find/Replace

Transform

Split

Combine

Match

Distance

Column: Date

Generate Dates

Make Date

Use Date

Infill Missing Dates

Make Time

Use Time

Column: Reshape

Each dialogue in this menu changes the number of observations (rows, records) in the data, and hence writes the resulting data into a new data frame.

Column Summaries

This dialogue is for calculating summary statistics either for a set of columns, or for a single column grouped into subsets by the levels of a single factor. The results can either be for all levels of a factor, or for a single specified level.

Use the Tabulation dialogues if you need summaries over more than one factor

The user chooses the summary statistics to display, which can include proportions or percentiles.

The results can be displayed in the output window. They can also be stored in columns of new data frame and therefore used in further calculations.

Examples are in the Introductory Guide, Chapter 12.3

R command:

General Summaries

Sub-Dialogue

Stack

This takes data in multiple short columns and stacks them into a single long column in a new data frame. This is usually because data have been entered as a "rectangle" rather than as columns. The Introductory Guide, Section 3.4 gives examples for experimental data and for monthly rainfall data, Section 3.5.
A factor column is added, that indicates to which original short column each observation belongs.
Optionally there are additional short (ID) columns, that need to be expanded to the length of the new column. This is the purpose of the ID option.

The dataset called experi2 illustrates the use of this dialogue.

Fig. 3.4a Experimental data - Standard layout of data for analysis by hand

	Block			
Treatment	1	2	3	4
1	330	288	295	313
2	372	340	343	341
3	359	337	373	302

Fig. 3.4b Layout for analysis

Count	Block	Treat
330	1	1
288	2	1
295	3	1
313	4	1
372	1	2
340	2	2
343	3	2
341	4	2
359	1	3
337	2	3
373	3	3
302	4	3

In Section 9.2 of the Introductory Guide we show how these data are stacked into the correct form for an analysis.

R command: melt

Unstack

This is used to change a single long column into multiple short ones. It is the converse of the Manage => Reshape => Stack dialogue.

The observations do not need to be in any particular order but you must specify a factor column that indicates to which new column each value is to be placed.

The carry option is for situations where you need to add information from further columns, so it can be used with the new shaped data.

The example below shows the dialogue and result of splitting the Count column from the experi.wor dataset into 3 columns, depending on the levels in the Treat column.

R command: dcast

Merge

Need to add help!

R command: merge

Append

Subset

Need to add help!

R command: subset

Random Subset

Need to add help!

R command: sample
replicate

Transpose

Need to add help!

R command: rbind.fill

Keys, Links and Notes

Add Key

View and Remove Keys

Add Link

View and Remove Links

Add Note

Data Object

Data Frame Metadata

Rename Data Frame

Need to add help!

R command:

Reorder Data Frames

Need to add help!

R command:

Copy Data Frame

Delete Data Frame

Need to add help!

R command:

View Metadata

Rename Metadata

Reorder Metadata

Delete Metadata

R Objects

View

Rename

Reorder

Delete

Export

Describe Menu

One Variable

Summarise

Graph

Frequencies

Rating Data

Two variables

Summarise

Graph

Frequencies

Correlate

Three or Four Variables

Summarise

Graph

Frequencies

Specific

Frequency Tables

Displays one-way, two-way, or multi-way tables that contain counts or percentages. The columns must have been defined as factors. Examples are given in the Introductory Guide Chapter 13.2.

Percentages can also be given, either overall or of any of the factors that define the table. The table can be presented either with, or without the margins.

A column of weights can be used and/or a filter column can be applied.

The resulting table opens in a new window. The menus change and further dialogues are available to manipulate and present the information in the table.

The simplest case of a one-way table is sometimes called tallying and can also be given using the Statistics ⇒ Summary ⇒ Group dialogue.

The Statistics ⇒ Tables ⇒ General dialogue permits further options, such as the display of counts and percentages together.

Any missing values in the defining factors are omitted from the table, and reported in the output window.
It is also possible to include summary statistics in tables as is shown in the dialogue Statistics ⇒

Summary ⇒ Column Statistics. For tables of counts the Statistics ⇒ Simple Models ⇒ Chi-square Test can then be used.

R command:

Summary Tables

Displays one-way, two-way, and multi-way tables. The columns defining the table must be factors. Examples are in the Introductory Guide, Chapter 13.3.

The table contains summary values for a variate. The user chooses which statistics to present, including the count, mean, maximum, minimum, sum and standard deviation.

A single table can display the results for more than one variate and for more than one summary statistic.

Choose the one-way table option to explore the data. This operates immediately, i.e. before the OK is pressed. This display may be "dynamic", in which case it changes as the dialogue is changed. Or it may be "static", in which case you may expand it to the whole screen. Lots will change - at least from here.

In the one-way display click on the 0,1,2 numbers at the top left of the display, shown below, to collapse the table. Alternatively, clicking on a + opens up the table to show the data that make up the summary. This is sometimes called "drill-down". It is useful as an exploratory tool to check on oddities in the data.

One-way display illustrating drill-down to show individual values

In a given table, the set of different statistics is considered like another factor. In the example there would be 3 "levels" to this factor. The only difference is that this factor does not have margins.

Clicking on the OK, produces the two-way table in its own window. This can then be edited. You can change the factors in the rows and columns, and exclude statistics, or levels of the factors. You can also merge levels, or change their order in the display.

You can also alter the format of the numbers in the display or the display of the cells themselves. You can also use "flooding" to display the results in a form that resembles a horizontal bar chart.

Where there are missing values in any variate then that row is omitted from the data.

See the Statistics ⇒ Tables ⇒ General dialogue if you need options that are not available from this dialogue.
For one-way tables the Statistics ⇒ Summary ⇒ Column Statistics may be used.

R command:

Multiple Response

This dialogue gives counts and percentages for multiple response data. The columns to be tabulated must previously have been defined as factors.
To use the dialogue, first choose which of the 3 layouts, shown below, correspond to the data. The terms such as multiple dichotomy are explained in the Introductory Guide, Chapter 13.8, which also includes examples of the resulting tables.

For the first 2 options above, namely Multiple Dichotomies and Multiple Responses, select the columns that correspond to these responses. Then select one or more ordinary factor columns that are also to make up

the table.
For the Separate table layout choose a single column (at the response level) and also a linking column that specifies which rows (or respondents) correspond to each response.
Then, as in the simpler Statistics => Tables => Frequency dialogue, you have the choice between tables of counts and percentages. (The complication here is whether you want counts of responses or respondents.)

The simplest table is a count of all the responses. Optionally, you may also display the total number of respondents, and/or the number who gave some response. The example, from Section 13.8, shows there were 11 responses, from 8 people, 6 of who responded.
Percentages: When specifying percentages you have the choice (as in the frequencies dialogue) between overall percentages, the percentages of each multiple response or the percentages of one of the other factors.

100%: The essential feature of the multiple response dialogue is to choose what constitutes 100%. This may be the number of responses, or respondents, or the number of active respondents, namely those that give at least one positive response. The second option is shown and all three options are in Section 13.8.

Unique responses: For the multiple responses and separate tables options, a further option is the "unique responses" checkbox. This is only relevant if the same respondent may give the same response more than once. For example if the response listed films seen recently, then a respondent may have seen multiple films, and also may also have seen the same film more than once. The multiple occurrences of the same response (e.g. same film) would be ignored if the unique responses checkbox is ticked.

Weights/filters: Weights and/or filters may also be applied. With the third "separate tables" layout there may be weights at either, or both levels.

Missing values: Missing values in the ordinary factor columns are treated as in the other dialogues. In the multiple response column(s) they are treated differently. The observations are ignored, but the whole row is not omitted.

R command:

Scatterplot

Need to add help!

R command:

Line Plot

Histogram

Need to add help!

R command:

Boxplot

Need to add help!

R command:

Dotplot

Need to add help!

R command:

Rugplot

Bar Chart

Bar Chart From Summary

General

Column Summaries

Tabulation

This will all change. There may not be a dialogue like this. But this is a record of the flexibility of the tabulation in the original Instat.

This provides a flexible dialogue to specify the data that are to be summarised in tabular form. Examples are in the Introductory Guide, Chapter 13.7.

Normally the set of factors that define the table are chosen first. Then choose one or more variates if appropriate, and add the way in which you wish to summarise the data. Choose a weight column if you wish. Then use the Add Statistic button to add it to the summaries that will be produced. This generality permits you to have different statistics or weights as you wish.

We describe the different options in turn:

If statistics columns are included when "Counts" are added, then the counts are normally of the non-missing cases in these columns. The other options give the count of the missing values, or the total count of the factors.
The count of missing does not include any missing values where the defining factors also have missing values. These rows are eliminated first.

If no statistics columns have been specified, for this statistic, then "Counts" gives the frequencies for each combination of the factor levels. The missing option does not then apply. The count of missing factor levels is reported in the output window.

If statistics columns are omitted, then this "Percentage" option gives the same tables of percentages as can be given with the Frequency tables dialogue, though here it is possible to combine this summary with others, such as the counts, in the same table. Specify which factor, if any is to give the 100%.
When statistics columns are included, the table gives a percentage of the totals of that statistic. For example if monthly rainfall is tabulated, this could give the percentage of the annual in different months. Again these percentages could be overall, or over a particular factor.
The remaining options, shown below, all require one or more summary columns to be specified. With "Totals", "Averages", "Extremes" or "Spread", you then choose which particular summary you require.

The remaining two options allow percentiles or proportions of the summary columns to be given.

Percentiles gives the values corresponding to the specified percentage point. For example, the 50% point is the median.
Proportions are the inverse. Here the value is given and the table then contains the proportion. The may be expressed as a percentage, and you may choose to give the proportion less than the value specified, or greater than or equal to the value

The figures above illustrate the statistics that can be used. The counts and percentages, give all that is provided by the Tables => Frequency, while other options include all the statistics provided by the Tables => Summary dialogue.
Once the table is produced, further dialogues are used to present the information as is needed.
Missing values in any of the factor columns, results in those rows being ignored. A report is given in the output window. If there are missing values in any of the variates that make up a table, then that row is also ignored for the whole table.
You may apply a filter column. With the filter column all rows with zero are taken as "false" and omitted. Non-zero values are taken as "true" and the rows are included. With the weight columns the rows with zero values are again omitted, while the non-zero values are used as the weight.

A further difference between filter and weight columns is that a filter is applied to the whole table, while weights are applied to an individual statistic. So you may choose different weight columns, or not to apply weights to some of the statistics in a table. For example, if you wish to ignore rows for a single variate only, then use a weight column with zeros in the corresponding rows, when specifying the statistics for that variate.

R command:

Graphics

Plot options

Facets

Need to add help!

R command:

Layers

Need to add help!

R command:

Titles

Axes

Need to add help!

R command:

Themes

Need to add help!

R command:

Multivariate

Correlations

This gives the simple (product-moment) correlation if there are 2 columns, and the correlation matrix if there are more than 2.
If there are just 2 columns, then the correlation coefficient can be saved, and you can also provide a confidence interval for the correlation. The Introductory Guide, Chapter 17.2 gives more explanation.

If you require the rank correlation, then use Manage => Transform first to give the columns of ranks. Might do it in this dialogue now!

R command: cor
cor.test

Principal Components

R command:

Canonical Correlations

R command:

More Graphs

Windrose

Need to add help!

R command:

Cumulative Distribution

Need to add help!

R command:

Population Pyramid

Inventory Plot

Use Graph

Combine Graphs

Themes

Export Graph as Image

View Graph

Model Menu

Can make some general points here perhaps?

Like having a Subset button.
And the limitation (?) that you first choose a data frame.
And the idea of the "shape" of the data.

General Points

Probability Distributions

Show Model

This dialogue is used to give probabilities and percentage points for a range of distributions. Examples are in the Introductory Guide Chapter 14.

The results can be for a single value, or for a set of values. The results can be saved into the worksheet, hence allowing the dialogue to be used to generate sets of tables.

The option to calculate expected frequencies is used in conjunction with the Statistics => Simple Models => Goodness of Fit dialogue, to compare observed and expected frequencies using a chi-square test. An example is in Case Study 3.
The option to plot distributions is described under Teaching and in the Introductory Guide, Chapter 14.

R commands:

Compare Models

Random Samples (Use Model)

Used to generate simulated random samples from different distributions. Multiple columns of data may be generated.
The parameters of the distribution must be given.
The sample size is the same as the current data frame.
To generate samples of a particular length (not necessarily the length of any existing data frame), then Use File ⇒ New and generate a new data frame of the required number of rows and possibly no columns.

Use a starting seed if you wish to repeat the same sequence, (perhaps for different distributions) on different occasions when you use the dialogue.

R command:
RNGkind
replicate

One variable

Gives confidence intervals and significance tests for data sampled from a single distribution. See the Introductory Guide, Chapter 15.2 for further information.
Use the Data drop down list to decide on the structure of your data. Usually the data will be a single column. You can use just part of a column by specifying the level of a factor in a second column. Check with the new Subset button.

Fit Model

Compare Models

Might be more general here - not just chi-square

This dialogue covers the use of the chi-square test for the goodness of fit of models. It is normally preceded by the Statistics => Probability Distributions dialogue, which includes an option to calculate expected frequencies.

Chi-square tests are not the only way to test goodness of fit and the Introductory Guide also considers the Kolmogorov-Smirnov test and the use of probability plots.

One special case of probability plots is for the gamma distribution and this is described in Chapter 11 of the Climatic Guide.

R command:

Use Model

Two variables

Fit Model

Choose Model

Use Model

Three Variables

Fit Model

Choose Model

Use Model

Four Variables

Fit Model

Choose Model

Use Model

General

Fit Model

Choose Model

Use Model

Other (One Variable)

Exact Results

Copied over from previous help and from first version of this help

Examples of results: (Capture new results later)

Normal model, one sample Column Yield Sample size 36 Minimum 19.1 Maximum 62.1 Range 43 Mean 40.581 Std. deviation 11.898 Standard error of mean = 1.983 with 35 d.f. 95% CI for mean 36.555 to 44.606

Use the Analysis drop-down list to specify whether you want a confidence interval for the mean or the variance.

The default is to give a Confidence Interval for the mean, but a Significance Test can be specified.

For the t distribution a further option is to Estimate % Points other than the mean, together with confidence intervals.

Probability plots can be produced as well as plots to show the data together with the estimated proportion and confidence interval. As well as the mean, the Confidence limits plots can also include the Prediction and Summary limits or the confidence limits for specific percentage points.

See the Two-samples dialogue for two groups and the One-way ANOVA dialogue for more than two.
If your data are clearly from a non-normal distribution, then you could look for the corresponding non-parametric method. However most of the inferences here for the mean are robust to departures from the normal distribution. So do not leap too quickly, without checking first in the Inference booklet, Section 9.

Also:
Proportions:

An example is in the Introductory Guide, Section 15.4.
Gives confidence intervals and significance tests for a proportion.
The data can be given in various ways, dictated by the setting in the layout drop down list. The simplest is to specify the column to be analysed. If this column contains just two different values, e.g.
1 1 1 0 0 1 0
then the larger will be taken as the "success". Where there are more than two, the user must specify what constitutes a "success". This may be a single value, a set of values, or all values less than (or greater than) a specified threshold.
You may analyse data from part of a column by specifying a level from a factor column.
Finally you may give the number of trials and the number of successes. The default is to give the estimated proportion of successes, together with a confidence interval for the proportion. The exact interval is given by default, but you may specify a normal approximation, or the simple normal approximation. More details are in the Teaching Ideas.
You may also add a significance test.
Plots are available to show the data together with the estimated proportion and confidence interval.

And:

Poisson

Gives confidence intervals and significance tests for a column of counts that are assumed to follow a Poisson model.
The data can be given in various ways, dictated by the setting in the layout drop down list. The simplest is to specify the column that is to be analysed.
You may analyse data from part of a column by specifying a level from a factor column.
You may give the summary statistics, i.e. the sample size, n, and the mean.
The final option is when the data are frequencies. In this case it is assumed that these are frequencies of values from 0 upwards.
The default output is to give the estimated proportion of successes, together with a confidence interval for the proportion. The exact interval is given by default, but you may specify a normal approximation, or the simple normal approximation. More details are in the Teaching Ideas.
You may also add a significance test. Plots are available to show the data together with the mean and confidence interval.

R command: t.test, z.test, var.test

Summary Data

Need to add help. Used to be combined with one-sample normal. Now different.

R command:

Non-parametric

A range of facilities for non-parametric analysis is provided, for situations up to 2-way ANOVA. In addition the chi-square test, under Statistics => Simple Models is often taught as part of non-parametric statistics.

See the Inference Guide for a summary of our views on non-parametric modelling.

This possibly soon to be incorporated into "Exact Results"

The non-parametric equivalent to the Statistics => Simple Models => One Sample and Two Sample.

The layout of the data is specified first. The one-sample tests normally perform a Wilcoxon test, though the simpler sign test can be specified.

For a 2-sample test the data can either be in 2 separate columns or in a single column, indexed by a factor.

An example of each test is in the Introductory Guide, Chapter 18.4.

R command: SIGN.test
wilcox.test

Other (Two Variables)

Two Samples

Soon to be incorporated into General

When Normal selected, then...

Gives confidence intervals and significance tests to compare data sampled from two normal distributions. Further information is in the Introductory Guide, Chapter 15.3.

For the analysis you may estimate the difference between two means or the ratio of two variances. For the two means situation your data may be paired or unpaired. If your samples are of a different size then it is obviously unpaired. Otherwise, see the Inference booklet, Section 5.4 if you are unsure.
The default is to give a confidence interval for the difference between two means or the ratio of the variances on the grounds that this is usually more informative than a significance test. A significance test can be specified.
Plots are available to show the data together with the estimated proportion and confidence interval.
The two-sample (unpaired) case extends to more than 2 samples with the Statistics => Analysis f Variance

=> One-Way dialogue. The paired case extends to the Statistics => Analysis of Variance => Orthogonal dialogue.
Following the ANOVA dialogue, this dialogue may then be used to give confidence intervals for the difference between any two means. This is one of the reasons for providing the options with the summary statistics. In this case, the residual degrees of freedom from the ANOVA are then also specified to ensure the correct degrees of freedom are used.

For proportions

Gives confidence intervals and significance tests to compare two proportions.

If the data columns just contain two distinct values, then the larger will be assumed to be a "success". Otherwise you must specify the criterion for success.

The default is to give a confidence interval for the difference in the two proportions. Optionally you may also test whether these two proportions could be equal.

Where the data are in two columns that are of the same length, a further option is to specify that they are paired. This is for situations where the same subject responded on two occasions, or to two questions.

Plots may be given. These show the original data, together with the confidence interval for the individual proportions and for the difference. You may specify the y scale for these plots.

Copied from old Instat where just Poisson

Gives confidence intervals and significance tests to compare two columns of counts, each assumed to follow a Poisson model.
In the dialogue use the Data drop-down list to specify the layout of your data. The data may be in two columns or in a single column with a second column being a factor column that defines the two samples. When the factor has more than two levels you need to specify which two will be compared.
The final option is to give the summary values, namely the sample size and mean for each sample.
The default is to give a confidence interval for the difference in the two means. Optionally you may also test whether the two means could be equal.
Plots may be given. These show the original data, together with the confidence interval for the individual means and for the difference.

Two Samples - Summary Data

The final option is just to give the summary statistics. Mention ANOVA situation.
The final option is to give the summary values, namely the sample size and number of successes for each sample.

R command:

Simple Regression

This is soon to be incorporated into "General" - above

This dialogue is for regression where there is just a single x-variable. In the equation y = a + bx, the "response variable" is the y. It is sometimes called the "dependent variable". The "explanatory" or "independent" variable is the x.

The results give the equation, together with a confidence interval for the slope of the line. You may choose to include the significance test of whether the slope of the fitted line could be zero (signifying no relationship) or another value that you specify. Optionally an ANOVA table to summarise the results can also be given.

The layout of the results is deliberately similar to that of the one-sample t test, to emphasise that the same concepts are being used. This is explained in the Introductory Guide Section 15.5.

A variety of graphs are possible and those concerned with teaching are considered under Teaching Ideas, with some examples in Section 15.5 of the Introductory Guide.

Section 17.3 shows more graphs, including a diagnostic plot of the residuals. In the regression dialogues the standard plots are always of the standardized residuals, and these are defined in Section 17.7. A number of statistics can be saved and these are also defined in Section 17.7.

Where the data include multiple observations at the same x values, it is possible to evaluate a "pure error" as illustrated in Section 17.8. This is a 2-step process using first an Analysis of Variance dialogue, and then the lack of fit option in the regression dialogue.

Details of the plotting options are given in Teaching Ideas.

R command: lm

One Way ANOVA

Soon to be incorporated into "Two Variables - General"

The one-way ANOVA extends the ideas of the 2-sample comparison (Statistics => Simple Models => Normal, Two Samples), to compare many groups or samples. The only layout of the data that is permitted is for the data for all the groups to be in a single column, with a second (factor) column indicating the group.

If the data for each group is in a different column, then use the Manage => Transformation => Stack dialogue first.
This dialogue can be used when there is an unequal number of observations in each group. If not, you may find the Statistics = > Analysis of Variance => Orthogonal gives more options for the analysis and a simpler layout of the results.
If there are unequal numbers in each group, and you need to investigate particular contrasts, then use Statistics => Analysis of Variance => General.

An example of the dialogue is in Chapter 16.2.

A wide variety of graphs may be given. Plots of the means can include the data points, or just be with the means. The means can optionally be connected as lines. Residual plots are the same for most of the ANOVA and regression dialogues.

The ANOVA plots are mainly to support teaching.

R command: aov

Non parametric Two Samples

The non-parametric equivalent to the Statistics => Simple Models => One Sample and Two Sample.

The layout of the data is specified first. The one-sample tests normally perform a Wilcoxon test, though the simpler sign test can be specified.

For a 2-sample test the data can either be in 2 separate columns or in a single column, indexed by a factor.

An example of each test is in the Introductory Guide, Chapter 18.4.

R command: SIGN.test
wilcox.test

Non-Parametric One-Way ANOVA

This is the non-parametric equivalent of the Statistics => Analysis of Variance => One Way dialogue. As with the ordinary ANOVA dialogue, the only layout of the data that is permitted is for the data for all the groups to be in a single column, with a second (factor) column indicating the group.

If the data for each group is in a different column, then use the Manage => Reshape => Stack dialogue first. The analysis is called the Kruskal Wallis test.

R command: kruskal.test

Other (Three Variables)

Simple Regression with Groups

This dialogue is for regression models where the model for y includes one variate, x, and one factor, f. Further information and an example are in the Introductory Guide, Chapter 17.4. It is assumed that a regression model is needed, use the Regression => Simple dialogue if necessary, to check this. The decision is then to choose which of the four possible models you require.

The alternative models are as follows:
1. A single line (i.e. the factor is not needed).
2. Lines with different slopes, but a common intercept.
3. Parallel lines, i.e. with multiple intercepts, but a single slope.
4. Separate lines.

The graphical display is optional and gives a plot for each model specified. It is usual to remove previous graphs (Shift+Ctrl+F3) before plotting. Then the plots can be tiled (F3) to compare the models. If you choose for the title to give summary statistics, it gives the residual sum of squares and the degrees of freedom for each plot.

When fitting you may choose relative or absolute estimates for the parameters. This does not affect the plots. We explain the reason for the option under Teaching Ideas.
Information on Residual plots is given here. Where???

R command: lm

Non-Parametric Two-Way ANOVA

This is the equivalent of a simple randomised block design, from Statistics => Analysis of Variance => Orthogonal or General.
The only layout of the data that is permitted is for the data for all the groups to be in a single column, with further columns indicating the levels of the block and treatment groups.

If the data are arranged in multiple columns, then use the Manage ⇒ Reshape ⇒ Stack dialogue first.

The analysis is called the Friedman test. An example is in the Introductory Guide, Chapter 18.4.

R command: friedman.test

Chi-square test

This isn't that bad here. It is in simple models, but different from the usual two sample case etc. It is more for data from surveys. It then leads into regression modelling (log-linear models) for more complicated situations.

Calculates the chi-square statistic for a contingency table. Further information and examples are in the Introductory Guide, Chapter 18.5.
The layout of the data can either be as a table, i.e. with the counts in multiple columns, or as a single column, indexed by two factor columns.
Sometimes the counts are calculated with the Statistics => Tables dialogues, and then saved back to the worksheet, together with the associated factor columns. In this case Instat has stored the corresponding factor columns when the table was created. Then this dialogue can be used for 2-way tables. For a more general method of fitting models to this type of data, see Statistics => Regression => Log Linear Models.

R command:

Other (General)

ANOVA

These are mainly for the analysis of experimental data. They can be used for a wide range of designs, see the Introductory Guide.

In this version there are also additional facilities for more general (unbalanced) designs.
There is an extensive system for fitting contrasts to look in more details at comparisons between treatment means.

Users looking for multiple comparison methods, such as Duncan's multiple range test should read our view of these methods in the Genanova Guide, the Inference Guide and in our Guide on the presentation of results.

This dialogue is used to analyse a wide range of orthogonal designs. These include the randomised complete block design, with one or more treatment factors, split plot and strip plot designs.

There is no limit on the number of factors that can be included. When there are multiple treatment factors, then first use the Manage => Data => Interaction dialogue to derive interaction columns, if these are to be included in the model.
The standard results include information on large residuals (that perhaps deserve further examination) and the full set of residuals can be saved and used in plots, etc.

You may also estimate contrasts for any of the treatment factors in the model. In the current version you must first enter the coefficients into columns in the worksheet. For example, if a factor has 3 levels you could enter
-2, 1, 1 as one contrast and 0, -1, 1 as a second.

With a split plot design the results include all the standard errors to compare means. You may have more than one split (e.g. split-split plot design), but then the results do not include all the standard errors of differences for the interaction tables.

Graphs can be given of the means, with, or without the data. Graphs of the residuals are as for all the ANOVA and regression dialogues.

For further information, see the Introductory Guide Section 16.4 for simple use, 16.5 for diagnostics, 16.6 for the analysis of a simple 3-factor experiment and also a 4-factor experiment that includes simple confounding. Section 16.7 considers the analysis of a split plot design, while 16.8 looks at how to examine the treatment means, and includes an interaction plot.

Coping with problems, such as missing values, is described in Section 16.9.

R command: aov

From old help:

Needs to change, because we no longer have this system. We may not have this dialogue, but let's see!

Despite the simplicity of the dialogue, the GENANOVA system provides powerful facilities for the analysis of experimental data. It is intended for possibly messy (non-orthogonal) designs that have a single blocking factor (or no blocking) and a single treatment factor. It can handle multi-factorial experiments through a powerful system of contrasts.
The display of the results is straightforward in simple trials. It gives the ANOVA table and treatment means, that may be adjusted if the design is non-orthogonal. Graphs of the (adjusted) means may be given, with, or without, the data points. Graphs of residuals are as for all the ANOVA and regression dialogues.
A simple example is in the Introductory Guide Chapter 16.3.
A powerful feature of the analysis is the production of user-defined contrasts. The typical display for a contrast is as follows:
Contrast Value se ssq F Prob >F Theta Efficiency
Trt_con 26.145 3.4026 1958.7 59.04 0.0000 0.349 0.907
Here the value of the contrast is the difference between two means. The contrast was (1, 0, -1) and is given with its standard error. The ssq, here 1959, with its associated f-value and probability, is to be interpreted as a proportion of the treatment sum of squares in the ANOVA table. Theta is the divisor in the contrast to give its standard error and the Efficiency is used mainly to assess the experimental design. See the Introductory Guide Section 16.10 for a complete example.
A special guide, based on this system, is available and is called The Statistical Background to ANOVA. It includes a description of a variety of further topics. For example, in some experiments the non-orthogonality may lead to disconnected parts in the experiment. This is reported in the output, giving the treatments that are in each part. The special guide, Section 5.2, gives more details

Regression

There are dialogues for simple regression (mainly for teaching) and for multiple regression.

For more details see the Introductory Guide. See the Climatic Guide for some examples of the use and misuse of correlation and regression.

This version also includes facilities for the analysis of log-linear models, which may be visualised as an example of regression modelling for non-normal data. Now much more general!

In the old Instat there are two headings and this is under multiple regression - and the next is general linear models.

The first step is to transfer all the columns (y and all x's) into the "terms box, that you might wish to include in the regression study. Then choose the y-variable.
Then transfer the terms needed for the current model and specify whether you wish to Add, Drop or Fit terms. If the Fit button is activated, this corresponds to a FIT command and fits the model specified. Once a model has been fitted, you can either choose to Fit a different model, or Add, or Drop terms from the existing model.
The results summarise the current model. They do not initially display the fitted equation on the grounds that it is not needed until you have decided on the model. The Estimates check box may be ticked whenever you need the equation. If necessary you can save the coefficients.
This is a dialogue that you might use repeatedly, in searching for the most appropriate model. Hence there is an Apply button. This is the same as OK, except it leaves the dialogue open.
An example of the process is in the Introductory Guide, Section 17.5.
The dialogue does not currently handle factor columns directly. If you need to include factors in the model then either use Statistics => Regression => General Linear Models or first use Manage => Data => Indicator.
For polynomial regression use Manage => Data => Polynomial first to calculate the necessary columns, see Section 17.9 for details.
Plots of residuals are as for all regression dialogues. Other topics in Chapter 17 of the Introductory Guide include regression diagnostics (Section 17.7), weighted regression and missing values (Section 17.10) and lack of fit (Section 17.8).

This dialogue extends the facilities for fitting multiple regression models, and is for models that include both factors and variates. If your model includes only variates, then it is simpler to use the Statistics => Regression = > Multiple Regressions dialogue.
You may choose any combinations of factors or variates to include in the terms that will form part of the model. Then specify the model and use either fit, or add or drop.
You may also choose to specify interaction terms. These are normally the interaction between 2 factors, or between a factor and a variate. For completeness we also allow quadratic and cross-product terms between variates.
When models are fitted that include factors, Instat automatically calculates the necessary indicator columns. The use of the details checkbox is described under Teaching Ideas. It is designed to show how Instat fits the models that include factors and interactions.
When fitting models involving factors the degrees of freedom associated with a factor are normally one less than the number of levels. The fitting options allow the user to specify whether the first or the last level will be omitted from the fitted model.
When interaction terms are fitted, the number of degrees of freedom for the interaction are adjusted automatically, depending on whether the main effects from that interaction have been included in the model.
An example of the use of this dialogue is in the Introductory Guide, Chapter 17.6. Graphs of residuals are as for all the regression dialogues.

R command: lm

Log-Linear

The y-variable should consist of counts. This is the same type of data as for a contingency table, see Statistics => Simple Models => Chi-square test, but here the counts must be in a single column, with other columns giving the factors.
Use Manage => Reshape => Stack first, if the data are in separate columns.
The model to be fitted is given by specifying the appropriate factors.
This is a dialogue that you might use repeatedly, in searching for the most appropriate model. Hence there is an Apply button. This is the same as OK, except it leaves the dialogue open.
Further information on the fitting of log-linear models is given in a special teaching guide. This guide also compares log-linear modeling with the use of simple chi-square tests and considered other topics, such as the handling of zeros in the data.
Examples of the fitting of log-linear models are in the Introductory Guide Chapter 18.6

R command:

Climatic Menu

Most of the options in this menu are designed to process daily climatic data. However there is generally nothing that is special to climatic data, hence the facilities may equally be used on daily records from other monitoring data, such as levels of sales, or pollution.

Similarly there is usually no check that the data are daily, hence the functions, such as spell-length calculations may also be useful for time-series data recorded on other scales, such as per minute, or per week.

The Manage option is a submenu to help setup the data for the rest of the options.

Most of the facilities in the Climatic menu are for daily data and Instat stores a single year of data in one column (Climatic Guide, Chapter 3). This column is usually of length 366, with a special dummy value for February 29th in non-leap years.

The Display Daily option activates Instat's DAY command and is intended for daily data, see Chapter 4 in the Climatic Guide. In contrast, the Summary dialogue (Chapter 5) makes use of Instat's STAtistics command, which is used also in the Statistics => Summary => Column Statistics dialogue.

The next 5 options are under the Climatic => Events menu and are for the Start of the Rains (Instat's RUN command, Chapter 6), for Spell lengths (SPEll command,Chapter 6) and for Water Balance calculations (also Chapter 7). Both the RUN command, that calculates running totals and the SPEll command can be used in many situations in addition to the summary of daily data.

The Evaporation facilities use the modified Penman method (Chapter 9), while the Crop Index is described in Chapter 10, and 12.2 for the reservoir option.

Chapter 8 in the Climatic Guide considers the analysis of temperature data, for which the calculation of Heat sums (degree days) is considered in Section 8.7.2.

Many other methods for processing climatic data are described in the Climatic Guide, Chapters 11 and 12, that are incorporated into the main statistics section in Instat. These include gamma distributions, correlations and regression ideas.

The fitting and use of Markov chain models is described under the submenu.

File

Open SST

Import Gridded Data (IRI)

ClimSoft

CliData

Export to CPT

Dates

Generate Dates

Make Date

Use Date

Infill Missing Dates

Make Time

Use Time

Define Climatic Data

Prepare

These dialogues are for the preliminary steps in setting up a worksheet for a climatic analysis, particularly for daily data.
A new worksheet can easily be created with the typical structure that is needed for Instat- where a year of daily data is in a single column (so a special code is needed for Feb 29th in non-leap years). Data can be imported into this worksheet from a Clicom system.

Then extra columns can be generated to support data summary on a range of scales, and data, such as evaporation, that is not on a daily basis can be interpolated.

R command:

Transform

This dialogue is simply to transform the (daily) data in the ways that are used in the the following dialogues on the start of the rains, spell lengths, etc. It produces further columns of the same length as the original data.

For example many definitions of the start of the rains use the running 2, 3, 5 or 7 day sums of the rainfall data. So this dialogue calculates these running totals. They can also be useful when calculating extremes as we could then calculate, for example the extreme value of the 3-day totals each year, or each season.

Spell lengths can also be calculated for any element as can the water balance.

Climatic Summaries

Used to summarise a set of years with daily data to a pentad, weekly, 10 day, monthly or other basis. In addition to the daily climatic data, the worksheet should normally contain a factor column, usually generated using Climatic => Manage => Make a Factor. Chapter 5 in the Climatic Guide illustrates the use of this dialogue for processing rainfall data and Chapter 8 describes the analysis of temperature data.

You have a choice of summary statistic. Usually use the sum to give rainfall totals, and the mean, max or min to summarise temperatures or other variables.

If the dialogue is used without a factor, then the annual summary is given. If you give the annual summary, or the summary for an individual factor level (e.g. for one month), then a time-series or other graph can be given. This permits an initial exploration of trend etc in the data to be assessed. Examples are in the Climatic Guide, Section 5.3. The restrict option, together with the Count summary may be used to give the total number of rainy, or hot days, etc.

When summarising the data using a factor, you can choose how to orientate the results, either by year, or by period. You may also write the summary data into a new worksheet.

R command: :

Start

Used to present or summarise daily rainfall data in a way that corresponds to the start of the rainy season for many definitions. For example:
"The first occasion after 1st April with more than 20 mm within a 3 day period."

The results may be displayed in one of 3 ways. We describe Simple and All in more detail under teaching. Here we consider the routine use of the dialogue to give the first occasion that the specified condition was satisfied.

You have to specify the condition. This normally consists of the earliest day (1st April in the example above), the number of days over which to total the rainfall (3 in the example above) and the criterion for success (20mm of rain is used above.)

The result will be a single column with the first day number each year that satisfies the condition. You may optionally save a second column that gives the quantity that was satisfied on that day.

If you do not know what condition to use, then we strongly recommend that you start by using the Simple option.
Sometimes a more appropriate definition of the start is based on water balance, rather than simply on the rainfall total. For simple water balance models this is provided using Climatic => Water balance.

More complicated definitions include a dry spell as well as a rainfall total, for example the first occasion with more than 20 mm and no 7 day dry spell in the following 20 days.

You may occasionally want more than just the day number of the start, perhaps also the rainfall total that triggered the start. This is the "Save quantities into…" option from the dialogue.

R command:

Spell

Used to calculate spell lengths, usually for daily climatic data. These are often dry spells for rainfall data as described in the Climatic Guide, Chapter 6, but could alternatively be hot, or cold spells for temperatures or calm spells from wind-speed data.

Three types of result can be given, called unconditional, conditional or simple. We describe the use of simple under teaching below.

You must give the range of values that observations must satisfy, for them to be within the spell. When processing rainfall data to give dry spells we often use the range 0 to 0.85mm, which defines all days with less than 0.85mm as a dry day.
The results give the longest spell length within each period that is specified, as shown in Chapter 6.3. You must give the initial day and the length of each period. When there are many periods, the starting dates and lengths are first entered into columns in the worksheet.

R command:

Extremes

This dialogue is designed to produce extreme values from daily data. The extremes may be for rainfall, temperature, river flow or any other element. They are usually the maximum values, but may also be minimums (for temperatures, etc).
The output is the set extreme values, and may also include the day number and year of the extreme. The default is to use all the data, but the calculations may be restricted to a particular part of the year.

Instead of getting the maximum (or minimum) value in each period - usually a year, you may opt to output all values that exceed a defined threshold. For example, for daily rainfall data you may request all values that exceed 80 mm. In this case some years may provide multiple observations, while others give none.

Usually the analysis will be for single values, but you may opt to calculate the extreme totals or means over a set of consecutive days. For example, if the objective were to calculate to risk of rainfall exceeding 100mm in a 3-day period, then the extremes would be calculated from running 3-day totals.

When providing all the values that exceed a given threshold you may opt to ignore values that are very close together. For example with daily data as follows:
X1 X2 X3
Day Rain Rain3
1 0 0
2 65 65
3 40 105
4 35 140
5 0 75
6 50 85
7 0 50
If you give a threshold as 80mm in 3 days, then days 3, 4, and 6 all exceed this threshold. You may choose to give just the maximum in each consecutive "group", in which case, you would give 140mm on day 4 and 85mm on day 6. You could insist that maximums are from groups at least 2 days apart, in which case you would give the single value of 140mm on day 4.

An alternative output is to give all the values, but also to add a "group" column, that specifies how many groups of values there are. This is given as a factor column, so the Statistics => Summary => Column Statistics dialogue could then be used to give the maximums (or other summary values) in each group.

Notes: If you give a low threshold then you must ensure that the worksheet is long enough to hold all the extremes that are generated. If necessary use the Manage => Resize Worksheet dialogue.

Once you have the columns of extremes, they may be processed simply, as for any other climatic "events". Or a distribution can be fitted, using the Statistics => Simple Models => Gamma, or Simple Models => Extremes dialogues.

R command:

End

Calculates the water balance from rainfall and evaporation data.
For the evaporation, a single column is specified, which may be input or calculated using the Climatic => Evapotranspiration dialogue. Alternatively a single value, e.g. 5mm per day can be specified. The capacity of the soil must also be given, if not, then 60mm is used as the default.

Results either give the water balance each day - we call this "simple" and describe under teaching - or to give the first occasion that a condition is satisfied. The example shown is the first time the balance is empty after 1st September, which might be used as the date of the end of the season.

The Reference Guide gives information on the incorporation of runoff (Water Balance (extended use)) when there are large rainfalls, on models where the evaporation is reduced when the water balance is low and also on making the analysis conditional on particular conditions up to a given date in the year.

The Climatic => Crop => Water Satisfaction dialogue also includes a calculation of water balance, but adds a crop water requirement that depends on crop coefficients that dictate the crop water requirements at different stages in the growing season.

R command:

Quality Control

Inventory

Display Daily

Displays daily data in tabular or in graphical form, or both. Chapter 4, Section 4.2 in the Climatic Guide gives examples and discusses the use of this facility.

When displaying in tabular form you may include monthly totals or means etc. This also provides the overall mean or total for the year.

There are also various options to clarify the presentation. You can choose how many decimal places to include (1 is the default) and rainfall data is often clarified by coding all zero values, so the rainy days stand out in the table.

With graphs the needle plots are appropriate for rainfall data, while lines are normally used for temperatures and other continuous variables.

R command:

QC Procedures

Climdex

This dialogue implements the 27 climate change indices from the RClimDex software, using the implementation from the R package climdex.pcic.

Describe

Rainfall

Temperature

Wind Speed/Direction

Sunshine/Radiation

General

Climatic Graphs

PICSA

The Climatic Guide, Chapter 7.3, describes the use of this dialogue.

R command:

Rainfall

Temperatures

Crops

Seasonal Climate Forecast Support

Open SST

Export to CPT

Correlations

Principal Components

Canonical Correlations

Model

Extremes

Circular

Markov

Fit Model

Historic (from old Instat)

The Climatic Guide, Chapter 7.4, describes the use of this dialogue.

R command:

Evaporation

R command:

Site

Instat includes the facilities to calculate evapotranspiration using the modified Penman-Montieth equation. The Site dialogue is the first part of the calculation and is used to enter the information applicable to the site for the data to be analysed, for example height above sea level. The Climatic Guide, Chapter 9 shows the use of this and the Penman dialogue for monthly data using an example from Kayes in Mali, and also for daily data from Bohicon in Benin.

R command:

Penman

Instat includes the facilities to calculate evapotranspiration using the modified Penman-Montieth equation. The Site dialogue is the first part of the calculation and is used to enter the information applicable to the site for the data to be analysed, for example height above sea level.

The Penman dialogue is used to specify the columns containing the climatic data. The columns needed for the calculations are then derived and given special names, as described in the Climatic Guide.

The Climatic Guide, Chapter 9 shows the use of this and the Penman dialogue for monthly data using an example from Kayes in Mali, and also for daily data from Bohicon in Benin.

R command:

Crop

R command:

Crop Coefficients

This dialogue is to generate the daily crop coefficients from values given for each growth phase of the crop. The values for each growth phase are in a variety of sources, particularly the FAO publication "Irrigation and drainage paper 56", by R. G. Allen et al (1998). This is available online from www.fao.org/ag/agl/aglw.

What is usually given is the duration and value for each of four growth phases. For example a 125 day maize crop may have durations of 20, 35, 40 and 30 days for the initial, development, mid-season, and late stages of the crop.

The coefficients are given as 0.3, 1.2 and 0.6 for the initial, mid-season and end of the crop. The way this translates in the Instat dialogue is as follows:

Stage	Duration(days)	Coefficient

1	20	0.3
2	35	0.3
3	40	1.2
4	30	1.2
5	0	0.6

So the coefficients are given at the start of each crop phase. In this case therefore the coefficient remains at 0.3 for the first 20 days and rises to 1.2 over the next 35 days. It then remains at 1.2 for stage 3, and then drops to 0.6 for the 4th and final stage.

The default is to interpolate with straight lines. You may choose to use moving averages to smooth any changes in the values.

Notes: These coefficients have to be adjusted for different lengths of the same crop. An agro-meteorologist can advise on the revised durations and values.

R command:

Water Satisfaction Index

Calculates the crop monitoring index from Frere and Popov (1979).
The minimum information you need is rainfall data for one or more years, evaporation data and a set of crop coefficients.
The default is for the season to start on the first period with more than 30mm, but many options are possible.
The options when there is surplus water can be varied. A conditional analysis is also possible. This example is analysed further in the Climatic Guide.
The crop index can also be used as a measure of the effectiveness of an irrigation scheme.
Once the index is calculated each year, the resulting values are plotted and processed using the ordinary commands in the statistics menu. Other crop models may be run outside Instat and the results processed in the same way.

Irrigate

Climatic => Crop => Water Satisfaction => Irrigate

This dialogue is part of the Climatic => Crop => Water Satisfaction dialogue

R command:

Degree Days

Uses columns of daily maximum and minimum temperatures to calculate degree days (accumulated temperatures).

R command:

Markov Modelling

The system for Markov modelling in Instat is in 3 parts as described in the Climatic Guide, Chapter 13. First the data are prepared for the model fitting. Then the models are fitted, and finally they are used.

The case to consider this approach to the analysis of the climatic data is also given in the Climatic Guide. It follows a study that compared 5 different methods for modelling rainfall data and proposes this approach as a sixth method.

Counts/Totals

This is the first dialogue in the sequence of fitting Markov models to the daily rainfall data, though other types of data can be used. It is used to summarise the daily data to give the counts of rain days. Optionally summaries of the amounts can also be given.

We are fitting 2-state Markov chains, and the threshold distinguishes between the values in the different states. Fitting a first-order chain produces 4 columns of counts (popup showing some results) giving the number of days with 'dd' 'dr' 'rd' 'rr' (popup below to explain notation)

R command:

Prepare for modelling

Totals the data, following the Markov Modelling => Counts/Totals dialogue, usually to give 5 day totals. Also adds a column giving the day number in the year, that is needed for the modelling process.

The reason for this stage in the process is explained in Section 13.3 of the Climatic Guide, see for example Fig 13.3g for an example of a file following the summary process.

Following the summary, some initial calculations are performed and plots of the chance(s) of rain are given. This is to enable you to choose the order of Markov chain model to use in the fitting.

During this process of initial calculation you will occasionally see the message
"…" in the output screen. This occurs when there are no rainy days in the summarised period and Instat has to evaluate 0/0. It correctly sets the result to a missing value, so this does not affect the further analysis.

Following this summary step use the Model Probabilities dialogue to fit the Markov chain models to the chance(s) of rain.

R command:

Fit Probabilities

Used to fit models to the chance(s) of rain as described in the Climatic Guide Section 13.4, see for example Fig 13.4d.
This dialogue controls the running of the @marfit macro. It asks the user to decide on the appropriate complexity of the models. To start the modelling process, choose the number of harmonics - usually 2 is a good starting point. You cannot have more than 5 and will probably need 3 harmonics at sites where the rainfall pattern is bimodal.

When you have an initial model you are asked whether you wish to accept the model, or (A)dd, or (D)rop terms. If you respond with A or D, then the model is changed by one term. You may also respond with AA or DD, and two terms will be added or dropped. This enables the model to be changed by a full harmonic (both the sine and cosine term) in one step.

Following this stage proceed to model the rainfall amounts if you chose to summarise them when using the initial Summary step.

The Markov dialogues use Instat macros, rather than built-in commands. This allows the (knowledgeable) user to change the macros if the models are not appropriate. For example, the limit of 4 harmonics would require a change in the macros.

R command:

Fit Amounts

Used to fit models, based on the gamma distribution, to the rainfall amounts. Prior to using this dialogue you must summarise the rainfall data and choose to include a summary of the amounts.

When you have an initial model you are asked whether you wish to accept the model, or (A)dd, or (D)rop terms. If you respond with A or D, then the model is changed by one term. You may also respond with AA or DD, and two terms will be added or dropped. This enables the model to be changed by a full harmonic (both the sine and cosine term) in one step.
The modelling process is described in the Climatic Guide, see in particular Fig 13.4e.

R command:

Prepare for Simulation

This step is needed if you chose to summarise the data when using the Summary process earlier. In that case the model gives the fitted values on a 5 day basis, (if you chose the default of 5 days in the summary). This dialogue then produces the fitted values on a daily basis, because this is needed by the subsequent commands that make use of the fitted model.

R command:

Simulation

Used to simulate daily data from a Markov chain model that has been generated, either from the dialogues in the system within Instat, or using alternative software.

If from Instat, then the fitted model will have named columns in a special way that simplifies the use of this dialogue. Otherwise name the columns yourself or choose them within the dialogue.

You may choose to add the simulated data to the same file or to write a new file. A new file is needed if you wish to simulate many years of data.

Note Instat's limit of 127 columns is a minor problem here, though 100 years usually provides reasonable results.
The method is described in Section 13.5 of the Climatic Guide.

Note also the facility to specify the starting random number, a feature that can greatly improve the precision of any comparisons between sites, or between models (perhaps for different periods of the record) within a site. The use of this feature is discussed in the Climatic Guide at the end of section 13.5 (see Fig 13.5n and o).

R command:

Spell Lengths from Model

The dialogue uses the Markov chain model to calculate risks etc. of long dry spells. It is therefore the equivalent of using the Climatic => Events => Spells, followed by the Climatic => Process dialogues, if the direct method of analysis is used.

The use of this dialogue is described in the Climatic Guide, Chapter 13.5.

R command:

Totals from Model

The dialogue uses the Markov chain model to calculate percentage points and probabilities for the total rainfall (or the number of rainy days) in 5, 7 10 day periods. It is therefore the equivalent of using the Climatic => Summary, followed by the Climatic => Process dialogues, if the direct method of analysis is used.

The use of this dialogue is described in the Climatic Guide, Chapter 13.5.

R command:

Procurement

Define Procurement Data

Prepare

Standardise Country Names

Merge Additional Data

Choose Countries for Analysis

Use Award Date

Use Signature Date

Set Reference Level

Define Contract Value Categories

Recode Numeric into Quantiles

Describe

Define Outputs

Model

Fit Model

Define Red Flags

Calculate Corruption Risk Index

Tests and Checks

Tools Menu

Run R Code

Restart R

Check for Updates

Clear Output Window

Save Current Options

Load Options

Options

View Menu

Statistics packages involve many windows. In R-Instat these include:

- Current Data Frame window (spreadsheet view of the data)
- Output (results) window
- Log window
- Script Window
- Possibly command windows
- and possibly many graphics windows.

Here we provide the options for managing these windows. A little time establishing the settings you find most convenient will make R-Instat much easier to use.

For example, if you open a worksheet, and then use Tile Vertical, followed by Save Window Positions, then this becomes the default for the current session. You can then use File ⇒ Options, so future sessions start in this same way.

Data View

Output Window

Log

Script Window

Column Metadata

Data Frame Metadata

Windows Layout

Open Last Dialogue

Help menu

The General Help (Contents) is the place to learn about the background to Instat. It also allows access to all other parts of Instat's help. They include the 4 guides, namely Introductory, Climatic, Reference and Tutorial.

Getting Started option explains about the Windows version.

Then there is access to the Dialogue Help. It gives information on the role of each menu and dialogue. You can either start at the top, which is the entry point from the Help menu, or from the Help button in any of the dialogue boxes. That takes you to help on the specific dialogue.

The help on each dialogue includes jumps to the data sets that can illustrate the dialogue, the examples in Instat's documentation that use the dialogue and the command(s) that correspond to the dialogue. They can be accessed through the general help, or directly.

About R

R-Instat provides an easy way to start using R. Here we discuss how users usually start with R. We give some resources we find are useful for those who want to learn R. We also discuss how you could "graduate" from using R-Instat to either using R instead, or to combining your use of R with continued use of R-Instat.

See:

Usual Ways to Use R
Using R from R-Instat

Usual ways to use R

Using R from R-Instat

R packages and Commands

Here we provide a list of the main R packages that are used within R-Instat. Each has a manual in a standard format.

A list of the main R packages used by R-Instat. Most are used for their commands and a few because of the interesting data sets that they provide.

Base

apply

apply(x, margin=1, function, na.rm= TRUE)
x- subset of the columns

Margin- a vector giving the subscripts which the function will be applied over: 1 indicates rows and 2 indicates columns.
Function- mean, sum, min, max, sd and for length we need function(x) {length(x)}.

Can we do a user defined function in this dialogue?

Na.rm= TRUE- omits missing values

cut

cut(x, breaks=c(), labels=c(), right = TRUE)
x – numeric variable to convert into a factor.

breaks - either a numeric vector of two or more unique cut points or a single number (greater than or equal to 2) giving the number of intervals into which x is to be cut.

labels – labels for the levels of the resulting category. If the labels are not specified, the labels are "(a, b]" interval notation.

right – the default is TRUE, which means the intervals are closed on the right. To make the intervals to be closed on the left, set right to FALSE

Dialogue
The argument for "closed on", is right. By default, the cut function produces intervals that are closed on the right. So the radio button for right should be selected.

droplevels

To remove unused levels from a factor:

droplevels(x), i.e. just give the factor name

factor

Convert to a factor:

factor(x = character(), levels, labels = levels, exclude = NA, ordered = is.ordered(x), nmax = NA)

Label a factor (or change the label):

levels= c( ) within the factor( ) command

c( ) contains the labels, e.g. c("New","Old",Trad"). This is the default order if the column was a character column before, i.e. it is in alphabetical order.

Reorder the levels (labels) of a factor column.

levels= c( ) within the function factor( )

c( ) contains the re-ordered levels (labels), e.g. c("Trad","Old","New"), would re-order the previous levels above.

head

interaction

interaction(..., drop = FALSE, sep = ".", lex.order = FALSE)

...	the factors for which interaction is to be computed, or a single list giving those factors.
drop	If drop is TRUE, unused factor levels are dropped from the result. The default is to retain all factor levels.
sep	string to construct the new level labels by joining the constituent ones.
lex.order	logical indicating if the order of factor concatenation should be lexically ordered.

merge

order

rank

rank(x, na.last=TRUE, ties.method= “average”)
x – variable

na.last – for controlling the treatment of NAs. If TRUE, missing values in the data are put last, if FALSE, they are put first.

ties.method – a character string specifying how ties are treated. The default for ordering ties is average. The possible options are; average, first, last, random, max and min.

RNGkind

sample

To permute rows:

sample(x, size, replace= FALSE)

x – variable

size – the size must always be equal to the length of the data frame

replace – always FALSE, since we want to sample without replacement

Dialogue

If the set seed box is checked, we must put the set.seed() command before the sample( ) command.

seq

set.seed

sort

subset

tail

car

recode

recode() from car package

recode(x, recodes, as.factor.result, as.numeric.result=TRUE, levels)

BSDA

SIGN.test

circular

climdex.pcic

dummies

dummy

dummy(x)

To produce dummy variables with an x variate, multiply dummy() by the x variate from the data frame (dummy()*variate)

foreign

ggplot2

mosaic

pdist

plotDist

qdist

plyr

rbind.fill

prettyR

describe

describe.factor

describe.logical

describe.variate

rep_n_stack

stretch_df

valid.N

reshape2

dcast

melt

rio

rio calls itself "A Swiss-Army Knife for Data I/O". It attempts to call the most effective packages for importing and exporting data from a wide range of statistical and other packages.

Convert

Export

Import

rrefine

sdc

stats

aov

cor

cor.test

friedman.test

glm

kruskal.test

lm

poly

relevel

To change the reference level of a factor use relevel() in stats package:

relevel(x, ref=" "), for example relevel(Village,ref="3")

wilcox.test

stringr

Datasets

We describe briefly the different datasets used in the R-Instat user guides from the "old" Instat. In each case we refer to the places in the guides that have used the data or that give further information.

A to D

Accident

A set of data used to calculate a time series with trends and residuals using the moving average, statistics and plot commands in the Introductory Guide, Chapter 19.3. It consists of car driver accidents for the years 1970 - 1984.

Activity

A small artificial data set to illustrate multiple responses. It is a survey of just 8 people, 5 men and 3 women, who record the activities they did in the last week (walking, jogging, cycling). Some did all three while others did not do any. It is used primarily in the Introductory Guide, Chapter 9.6 to show the different ways in which this type of data can be organised and in Chapter 13.9 where different types of table are shown.

Baden

These data are from Mead, Curnow and Hasted (2002) and are frequencies of different combinations of eye colours that were recorded for 6800 males in Baden. The objective of the exercise was to investigate the association between hair and eye colours.

The data are used in the Introductory Guide, Chapter 18.5 to illustrate the chi-square test.

Beans

A fertilizer experiment on the yield of beans, which is confounded in blocks of size eight. The data are from Cochran and Cox [1957], pages 188-192 and are used in the Introductory Guide Chapter 16.6 of to illustrate the analysis of a confounded experiment.

Blackleg

These data are from Mead, Curnow and Hasted (2002) and are taken from a much larger set of data from a large factorial experiment. One object of the experiment was to examine the effects of different fertiliser treatments on the incidence of blackleg for potato seedlings. A number of seedlings were examined for each treatment and classified as contaminated by blackleg or free.

The data are used in the Introductory Guide, Chapter 18.5 to illustrate the chi-square test.

Bohicon

Daily climatic data from Bohicon, Benin - 1983.

X1 Day number
X2 Wind speed (m/s)
X3 Sunshine hours
X4 Vapour pressure (mb)
X5 Minimum RH (%)
X6 Maximum RH (%)
X7 Minimum temperature (Deg C)
X8 Maximum temperature (Deg C)

Bpress

Blood pressure data for 13 people. This is used in Case Study 9 - "Least squares without calculus".

Brazil

Data on rice yields and weather data. Used to discuss the value of regression methods in studies of crop weather relationships (Climatic Guide 11.6.2).

Chick

Data on the weights of 2 groups of chicks. This example is taken from "Statistical Methods in Agriculture and Experimental Biology" by Mead, Curnow and Hasted [2002]

It is described in more detail in Case Study 6 - "The t-test - bookwork plus", which is on the teaching of two-sample t-tests.

Compact

MSTATC is an old statistics package, though it still appears to be in use.

This is a copy of the two MSTATC files called Compact.txt and Compact.dat. These data are described on pages 9-17 of the MSTATC Guide. The data are from and experiment to study the effects of soil compaction on bean yields. The layout is a split plot design with 4 replicates. The two treatment factors are the soil compaction at 2 levels and bean variety at 15 levels. There are therefore 120 subplots in the trial.

There are 9 columns of data in the file. The first 3 columns give the Rep, Compaction and Entry (Bean Variety). The final column also gives the bean variety, but transformed into the integers 1 to 9, because MSTATC, (like the original Instat), cannot cope with factor levels other than positive integers from 1 upwards.

The remaining 5 columns are the pod length, yield, seeds per pod, 1000-seed weight, and the number of pods per plant.
The data may be imported into Instat using the File => Import/Export dialogue and then analysed using the balanced ANOVA dialogue. Check what to do now - presumably import and then save the resulting Instat or

R-Instat file.

The results shown in the MSTATC Guide are of the number of pods per plant. The 1000 seed weight is of some interest, because it contains 1 or 2 very odd observations. This shows clearly the importance of a critical analysis of the data, as described in the good practice guide on data analysis with MSTATC and Genstat.

Coral

These data are 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 the lagoon were being degraded, due to pollution and other human activities.

More information about the dataset can be found in the Introductory Guide, Chapter 3.8.

Cucumber

MSTATC is an old statistics package, though it still appears to be in use.

This is a copy of two MSTAT files, called Cucumber.txt and Cucumber.dat. The data are described in the MSTATC Guide page 9-1. There are 2 columns, with the first giving the population factor and the second giving the fruit number for 30 plants. There were 11 populations and 5 replicates in this trial. Once imported the last 11 observations should be deleted, as they contain the means.

The data may be imported into Instat using the File => Import/Export dialogue and then analysed using one of the Analysis of variance dialogues. Check what to do now.

This example is also used in the good practice guide on the analysis of experimental data using MSTATC and Genstat.

E to K

Experi

A simple randomised block experiment with 4 blocks and 3 treatments. It is used in Chapter 16 in the Introductory Guide to illustrate analysis of variance.

This example is also used to show how regression methods, rather than simple ANOVA, can be used to fit models to experimental data.

Data:

Count	Block	Treat
330	1	1
288	2	1
295	3	1
313	4	1

372	1	2
340	2	2
343	3	2
341	4	2
359	1	3
337	2	3
373	3	3
302	4	3

Experi2

A simple randomised block experiment with 4 blocks and 3 treatments. The Manage ⇒ Reshape ⇒ Stack dialogue can be used to get the data in Experi.

Data:

	Block			
Treatment	1	2	3	4
1	330	288	295	313
2	372	340	343	341
3	359	337	373	302

Heart

Data on the length of stay in hospitals of 202 patients who were admitted after heart attacks. Used in Case Study 11 - "To group or not to group" to show how raw data, rather than merely grouped frequencies should be used in teaching.

Household

A hierarchical dataset from 6 households, where 'House (X3) identifies the house and 'Person (X4) of length 20, identifies the household to which each individual belongs. X1 and X2 contain data relating to the household. This worksheet is used to illustrate the Manage => Transformations => Expand dialogue.

Check what to do here, because now different.

Huda

Data on the relationship between crop yield and weather data, from Agric Met 15 1975 71-86. Used to discuss the value of regression methods in studies of crop weather relationships (Climatic Guide, Chapter 11.6.1).

Data:

Insects

Data on the number of moths recorded in 82 traps. Used in Case Study 3 - "Testing goodness of fit".

Kayes

Monthly data from Kayes in Mali, used to illustrate the calculations of evapotranspiration (Climatic Guide, Chapter 9).

Kundrip

Clicom is an old data management system for climatic data. These data may be deleted, or used in a new tutorial for Climsoft?

These are 2 years of data from the imaginary station called Kundrip, that is used in the Clicom tutorial. They have been exported as shown in the Climatic Guide, Chapter 3, Fig. 3.9. The data consist of daily maximum temperatures (element 002) and minimums (element 003), for the years 1980 and 1981.

They have been exported by element, i.e. each record includes one month of data for a given element. They are ASCII files and may therefore be inspected with a simple editor, for example using Edit => Text within Instat. They can be imported into an Instat worksheet using the Clicom dialogue.

Kundrip_obs

Clicom is an old data management system for climatic data. These data may be deleted, or used in a new tutorial for Climsoft?

This is one year of data from the imaginary station called Kundrip, that is used in the Clicom tutorial. They have been exported by as shown in the Climatic Guide Chapter 3, Fig. 3.14. The data consist of the 4 elements, rainfall, max and min temperatures and sunshine hours for the year 1980.

They have been exported by observation, i.e. each record is a day. They may be imported into Instat using the Clicom dialogue, stating that the import is by observation.

Kurun7

This worksheet consists of weekly rainfall totals from Kurunegala, Sri Lanka, for 1950 to 1983. They are calculated from the daily data given in the Kurunegala data, and are used to show different methods of estimating percentage points of rainfall totals.

Kurunegala

This worksheet contains daily rainfall data for 34 years, from Kurunegala, Sri Lanka, for 1950 to 1983.

They are particularly used to illustrate the fitting of Markov chain models in the Climatic Guide, Chapter13.4, Fig 13.4j for a site where the rainfall pattern is bimodal (Fig. 13.5h) and also one where it is desirable to fit a second order chain (Fig. 15.5l). They show the fitting of gamma distributions to decadal rainfall totals where the full analysis, starting from the daily data, uses an Instat macro.

The analysis of extreme events, described in Section 11.4 of the Climatic Guide also uses these data, and they form the basis of the kurun7 data set, that is used extensively to show different methods of estimating percentage points of rainfall totals.

L to R

Lettuce

Data on the yield of lettuces to illustrate the analysis of a split plot experiment, taken from "Statistical Methods in Agriculture and Experimental Biology" by Mead, Curnow and Hasted [2003].

Madaw

Daily rainfall data from Madawachchiya, in the North of Sri Lanka, from 1891 to 1978. Used to illustrate the use of the CROp command with irrigation (Climatic Guide, Chapter 12.2.3).

Malawicr

This worksheet consists of 10 day data for 27 years from a site in Malawi. Details are as follows:

X1-X27 Rainfall totals for decades (from 1st Oct.) for 1954/5 to 1982/3 - omitting 1962/3 & 1981/2.
X28 Evaporation totals for decades (from 1st October)
X29 Crop coefficients for groundnut (11 periods)
X30 Crop coefficients for cotton (17 periods)
X31 Numbers 1,2,... 17 (for plotting)
X32 Numbers 1,2,... 24 (for plotting)

These data are used to describe a preliminary analysis of the rainfall data (Climatic Guide, Chapter 10.4.1) and then particularly to illustrate the use of the FAO crop index where many years of data are available.)

Niatemp

These data are monthly maximum and minimum temperatures from Niamey, Niger for a 7 year period. They are used to illustrate plotting in the Climatic Guide, Chapter 2, Fig. 2.11.

Ntemp

Daily maximum and minimum temperatures for Niamey, Niger from 1961 to 1980. Maximum day numbers are in x3-x22 with the corresponding minima are in x23-x42. X1 is the day number (1 - 366) and x2 is free and normally filled using Climatic => Manage => Make Factor, with the factor chosen for monthly or decadal summaries. This worksheet is extensively used in the Climatic Guide, Chapter 8.4.

Poverty

These data are from a survey of 1789 households in Malawi and is 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. More details can be found in the Introductory Guide, Chapter 3.7.

Rain

Not copied correctly. Copy from full help

Rain10

Not copied correctly. Copy from full help

RainMay

Not copied correctly. Copy from full help

RdTemp

Not copied correctly. Copy from full help

Regress

Not copied correctly. Copy from full help

Reptile

Not copied correctly. Copy from full help

Runoff

Not copied correctly. Copy from full help

S to W

Sama57

Not copied correctly. Copy from full help

Samaru56

Not copied correctly. Copy from full help

Samdis

Not copied correctly. Copy from full help

Sammonth

Not copied correctly. Copy from full help

Samrain

Not copied correctly. Copy from full help

Samsmall

This worksheet consists of a subset of the daily rainfall data data available for Samaru, Nigeria and is for the 11 years from 1930 to 1940. This example is used extensively in the Climatic Guide to illustrate the different methods of processing the rainfall records.

The analyses include the calculations (Chapter 5, Fig. 5.3) of statistics for monthly and decade totals (Fig.

5.9), that are presented both in tabular (Fig. 5.5) and graphical (Fig. 5.6) format.

Calculations involving the start of the rains are described in Chapter 6 and on dry spells and water balance in Chapter 7 of the climatic guide.

This example is also used to illustrate the method of fitting Markov chain models, described in the Climatic Guide, Chapter 13.

Storm

These data are a sample from a detailed hydrological data in Malawi. They consist of 5 minute records from 50 storms. They are described in detail and analysed in Chapter 12.3 (Case Study 3) in the Climatic Guide.

Sulphur

This dataset is the same example as the one used by the statistics package Genstat. It is used, as in Genstat, to demonstrate the wind rose.

Survey

This is R-Instat's all-purpose example. It gives yields and other variables from a rice survey of 36 farmers. It is the basis of the case study on surveys and is also used to illustrate the tabulation of survey data in the Introductory Guide, for both simple tables and tables of mean yields.

We also use it to illustrate analysis of variance (ANOVA) with unequal group size and to introduce ideas of the teaching of ANOVA.

It is also used extensively in the regression chapter in the Introductory Guide, on the use of regression models when factors are to be included (Chapter 17.6 and Chapter 17.7) and to illustrate problems of collinearity (Chapter 17.8).

Survival

The data are from Collett (2003), "Modelling Survival Data in Medical Research" and give 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 cancer. The data are used in the Introductory Guide, Chapter 19.2.

Thames

An example of a data set used in multiple regression. It is taken from "Statistical Methods in Agriculture and Experimental Biology" by Mead, Curnow and Hasted [2002]. 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.

Wafric2

Monthly rainfall totals for 28 years from 2 neighbouring stations in Benin (West Africa). They show the use of plots and correlations, including autocorrelations to study possible relationships between the successive months and stations.

Licence Information