

S-Plus for the Analysis of Biological Data

**Rhonda Jones
Robin Gilliver
Simon Robson
&
Will Edwards**



This manual was first produced in 2009. It was made possible by a Teaching and Learning Development Grant from James Cook University. The authors are grateful for the support provided by Dr Nick Szorenyi-Reischl, Director of Teaching and Learning Development.

© 2010

Contents

	Preface	xi
	Why S-Plus?	xi
	How to use the manual	xiii
	To the instructor	xiii
	Acknowledgements	xiv
	Typesetting conventions used in the manual	xiv
1	Introduction to S-Plus	1
1.1	Starting S-PLUS in Windows	2
	Choosing or setting a working directory	2
1.2	The S-Plus main program window	3
	Getting help	3
1.3	Import data to create a new S-Plus data set	4
1.4	Create a new empty data set to enter sample data manually	6
1.5	The Object Explorer	7
	Changing the properties of an object	8
	Examining the details of a Data object	8
1.6	Data types in S-Plus	9
1.7	Data transformation: creating and modifying data	10
	Calculating a time interval in days	11
	Calculating a time interval in weeks	12
	Calculating a proportional weight gain	12
	Calculating a mean weekly growth rate	12
	Creating a logical variable	13
	Single-function transforms	14
	Other options provided by the Data menu	14
	Special values and reserved words	14
1.8	Data Objects in S-Plus	14
1.9	Introduction to the Commands Window	15
	Using the commands window as a calculator	15
	Giving names to objects: assignment commands	16
	Creating vectors	17

- Working with vectors 19
- Using data frames on the command line 20
- Working with parts of a data frame 21
- Creating a data frame in the commands window 22
- 1.10 Using the Script Window 22
 - Opening the script window 22
 - Entering and running a script 22
 - Writing functions 23
- 1.11 S-Plus language and functions 28
- 1.12 References and further reading 28
- Test Your Skills 29

- 2. Displaying data 31
 - 2.1 Displaying frequency distributions 31
 - Bar graphs for categorical data 31
 - Frequency tables and histograms for numerical data 34
 - 2.2 Quantiles of a frequency distribution 36
 - Plotting a cumulative frequency distribution 36
 - 2.3 Associations between categorical variables 36
 - Creating a grouped bar plot 37
 - Creating a stacked bar plot 38
 - Creating a mosaic plot 39
 - 2.4 Comparing numerical variables between groups 40
 - Using box plots 41
 - Using trellis graphics to compare histograms 41
 - Comparing cumulative frequencies for different groups 43
 - 2.5 Displaying relationships between a pair of numerical variables 44
 - Scatter plots 44
 - Line graphs 45
 - Putting several graphs on the same graph sheet 45
 - Varying symbols between groups on the same plot 46
 - Plotting fitted lines to scatter plots 47
 - Test your skills 50

- 3. Describing data 55
 - 3.1 Arithmetic mean and standard deviation 55
 - Data as individual values 55
 - Data as a frequency table 56

-
- 3.2 Median and interquartile range 57
 - 3.3 How measures of location and spread compare 58
 - Descriptive statistics with the GUI 58
 - 3.4 Proportions 61
 - Calculating proportions using the GUI 61
 - Calculating proportions using the commands window 61
 - Test Your Skills 63

 - 4. Estimating with uncertainty 65
 - 4.1 The sampling distribution of an estimate 65
 - 4.2 Measuring the uncertainty of an estimate 68
 - The standard error of the mean 69
 - 4.3 Standard errors and confidence intervals for the sample mean from the GUI 69
 - Test your skills 71

 - 5. Probability distributions 73
 - 5.1 Some terminology 73
 - 5.2 Probability 73
 - 5.3 What is a probability distribution? 74
 - 5.4 Using S-Plus to calculate probabilities for a binomial distribution 75
 - Calculating binomial probabilities using the GUI 75
 - Calculating binomial probabilities using the commands window 77
 - 5.5 What other information might you need from a probability distribution? 77
 - 5.6 Another common discrete probability distribution: the Poisson 78
 - 5.7 Continuous probability distributions in S-Plus: the normal distribution 79
 - 5.8 Other key continuous probability distributions 82
 - The Chi-square distribution 83
 - The *t*-distribution 84
 - The *F*-distribution 85
 - Test your skills 85

 - 6. Hypothesis testing In preparation

 - 7. Analyzing proportions 89
 - 7.1 The binomial distribution 89
 - Calculating binomial probabilities using the commands window 90
 - Properties of the sampling distribution for a proportion 91
 - 7.2 Testing a proportion: the binomial test 92

- 7.3 Estimating proportions 93
 - Estimating the standard error of a proportion 93
 - Estimating confidence limits for a proportion 93
 - Test your skills 95

- 8. Fitting probability models to frequency data 97
 - 8.1 Example of a random model: the proportional model 97
 - 8.2 χ^2 goodness-of-fit test 98
 - 8.3 Assumptions of the χ^2 goodness-of-fit test 99
 - 8.4 Goodness-of-fit tests when there are only two categories 100
 - 8.5 Fitting the binomial distribution 101
 - Testing goodness of fit to a binomial using frequency data 101
 - Testing goodness of fit to a binomial distribution with individual values 102
 - 8.6 Random in space or time: the Poisson distribution 103
 - Using the S-Plus GUI to test goodness-of-fit to a Poisson distribution 104
 - Using the commands window to test goodness-of-fit to a Poisson distribution 108
 - Test your skills 110

- 9. Contingency analysis: associations between categorical variables 113
 - 9.1 Associating two categorical variables 113
 - Contingency tables, proportional plots, and a χ^2 contingency test on categorical data for individuals 113
 - Creating a mosaic or stacked bar plot 115
 - 9.2 Estimating association in 2 x 2 tables: odds ratio 116
 - Using S-Plus to calculate odds and the odds ratio 116
 - 9.3 The χ^2 contingency test for n x n tables 117
 - What if S-Plus warns that some expected values are too low? 119
 - 9.4 Fisher's exact test 121
 - 9.5 Log-linear models and G-tests 122
 - Two categorical variables: using log-linear modelling to execute a G-test 122
 - A more complex example 124
 - Test your skills 127

- 10. The normal distribution 129
 - 10.1 Bell-shaped curves and the normal distribution 129
 - 10.2 Exact probability estimates for normal distributions 130
 - 10.3 Properties of the normal distribution 131

-
- 10.4 The standard normal distribution 132
 - Using normal distributions to answer questions about populations 133
 - 10.5 The normal distribution of sample means 134
 - 10.6 The Central Limit Theorem 135
 - 10.7 The normal approximation for the binomial distribution 135
 - Using the normal approximation for the binomial 136
 - Using the binomial probabilities 136
 - Test Your Skills 138

 - 11. Inference for a normal population 139
 - 11.1 The t -distribution for sample means 139
 - Using S-Plus to find values of t from probability values 141
 - Using S-Plus to find probabilities from the values of t 142
 - 11.2 The confidence interval for the mean of a normal distribution 143
 - Calculating confidence limits from the original data in the commands window 144
 - 11.3 The one-sample t -test 145
 - The effects of larger sample size: body temperature revisited 147
 - 11.4 Confidence intervals for the standard deviation and variance of a normal population 148
 - Test Your Skills 150

 - 12. Comparing two means 153
 - 12.1 Paired samples versus independent samples 153
 - 12.2 Paired comparison of means 154
 - 12.3 Two-sample comparison of means 156
 - A two-sample t -test where variances can be assumed to be equal 156
 - A two-sample t -test where variances cannot be assumed equal 158
 - 12.4 Using the correct sampling units 159
 - 12.5 Avoid indirect comparisons 160
 - 12.6 Interpreting overlap of confidence intervals 161
 - 12.7 Comparing variances 161
 - Test your skills 163

 - 13. Handling violations of assumptions 165
 - 13.1 Detecting deviations from normality 165
 - Graphical methods 165
 - Formal tests of normality 168

-
- 13.2 When to ignore violations of assumptions 168
 - 13.3 Data transformations 168
 - 13.4 Nonparametric alternatives to one-sample and paired *t*-tests 170
 - The sign test 171
 - 13.5 Comparing two groups: the Wilcoxon rank-sum test (Mann-Whitney U-test) 172
 - Test your skills 175

 - 14. Designing experiments In preparation

 - 15. Comparing means of more than two groups 179
 - 15.1 The analysis of variance 180
 - Executing a one-way ANOVA in S-Plus 181
 - Interpreting and reporting the results: the formula 182
 - Interpreting and reporting the results: sums of squares, degrees of freedom, and mean squares 183
 - Interpreting and reporting the results: the ANOVA table 184
 - Interpreting and reporting the results: the R^2 value 184
 - Interpreting and reporting the results: summarizing the data values 185
 - What you should report 185
 - How sums of squares are calculated 185
 - 15.2 Assumptions and alternatives 187
 - 15.3 Planned comparisons 188
 - Planned comparisons between two means 189
 - 15.4 Unplanned comparisons 190
 - Testing all pairs of means 190
 - 15.5 Fixed and random effects 192
 - 15.6 ANOVA with randomly chosen groups 192
 - Variance components and repeatability 194
 - Test your skills 195

 - 16. Correlation between numerical variables 197
 - 16.1 Estimating a linear correlation coefficient 197
 - The correlation coefficient 198
 - Standard error and confidence interval 199
 - 16.2 Testing the null hypothesis of zero correlation 199
 - Reporting the results of Pearson's correlation 200
 - 16.3 Assumptions 201
 - 16.4 The correlation coefficient depends on the range 201

-
- 16.5 Spearman's rank correlation 203
 - Reporting the results of Spearman's rank correlation 204
 - Assumptions of Spearman's rank correlation 205
 - 16.6 The effects of measurement error on correlation 205
 - Test your skills 206
 - 17. Regression 209
 - 17.1 Linear regression 210
 - The method of least squares 211
 - Executing a regression analysis through the Statistics menu 211
 - Interpreting and reporting the output: the function call 213
 - Interpreting and reporting the output: the residuals summary 213
 - Interpreting and reporting the output: coefficients and their standard errors 214
 - 17.2 Confidence in predictions 214
 - Interpreting and reporting the output: the prediction interval 215
 - Interpreting and reporting the output: confidence bands 216
 - Interpreting and reporting the output: the R^2 value 217
 - Interpreting and reporting the output: testing hypotheses about the regression line 217
 - 17.3 Doing the analysis in the commands window 218
 - 17.4 What you should report 218
 - 17.5 Assumptions of regression 219
 - Outliers 219
 - Detecting deviations from the assumptions: linearity 221
 - Detecting variations from the assumptions: non-normality and unequal variance 223
 - 17.6 Transformations 224
 - 17.7 The effects of measurement error 225
 - 17.8 Non-linear regression 225
 - A curve with an asymptote 225
 - Quadratic and polynomial curves 227
 - Formula-free curve fitting 228
 - Logistic regression: fitting a binary response variable 229
 - Test your skills 233
 - 18. Multiple explanatory variables 237
 - 18.1 Defining a model in S-Plus 238
 - 18.2 Analyzing experiments with blocking: the randomized block design 239
 - Analyzing data from a randomized block design 239

18.3	Analyzing factorial designs	242
	Analysis of two fixed factors	242
	Reporting the results of a factorial ANOVA	245
	Handling unbalanced factorial ANOVA designs	246
18.4	Adjusting for the effects of a covariate	247
18.5	Nested analysis of variance	249
	Executing a mixed-effects ANOVA via the Mixed Effects dialogue	251
	Analyzing a mixed-effects model by recalculating F-values and probabilities from a fixed-effects analysis	253
18.6	Assumptions of linear models	254
	Test your skills	256
19.	Computer-intensive methods	In preparation
20.	Likelihood	In preparation
21.	Meta-analysis: combining information from multiple studies	In preparation
	<i>Appendix 1: Working with the command line</i>	265
	Chapter 1: A beginning collection of useful functions	265
	Functions to create data objects	265
	Functions for basic descriptive statistics	266
	Functions to test or change data type	267
	Function to create or modify the ordering of factors	268
	Functions to aggregate and group data	268
	Functions to inspect variables	268
	Functions associated with probability distributions	269
	Basic mathematical functions	269
	Chapter 2: Functions for plotting data	270
2.1	Displaying frequency distributions	270
	Bar graphs and dot plots for categorical data	270
	Frequency histogram	271
2.2	Cumulative frequency distribution	272
2.3	Associations between categorical variables	272
	Mosaic plot	272
	Grouped bar plots	272
	Stacked bar plot	272
2.4	Comparing numerical variables between groups	273

	Trellis graphics	273
	Comparing frequency histograms	274
	Boxplots	275
2.5	Displaying relationships between a pair of numerical variables	275
	Scatter plots	275
	Line graphs	275
	Varying symbols between groups on the same plot	276
	Plotting fitted lines to scatter plots	276
Chapter 3: Functions for describing data 278		
3.1	Examining the whole data frame	278
3.2	Descriptive statistics for individual vectors	279
	Measures of location	280
	Measures of dispersion	280
	Measures of distribution shape	281
	To calculate descriptive statistics for subsets of an individual vector	281
Chapter 4: Estimating with uncertainty – functions for calculating standard errors and confidence limits 282		
4.1	Standard error of the mean	282
4.2	Confidence limits of the mean for normally distributed data	283
4.3	Confidence limits for the variance and standard deviation for normally distributed data	284
4.4	Confidence limits for descriptive statistics which do not require the assumption of normality	284
Appendix 2: Scripts used in each chapter 286		
Chapter 1:		
	plot.summarize()	
	Generates a set of descriptive plots and returns summary statistics for a numeric vector	286
	growth.rate()	
	Calculates and returns growth rate per time unit given start and end sizes, and times	286
Chapter 4:		
	sample.means()	
	Calculates and returns the means of a set of random samples taken from a numerical vector x	287
Chapter 7:		
	CI.p.agresti()	
	Calculates and returns a proportion and its confidence limits using the Agresti-Coull approximation for confidence limits	287

CI.p.exact()

Calculates proportion and its confidence interval with a specified tolerance 288

Chapter 8:

chisquare.gof()

Executes a chi-square goodness of fit for any specified set of observed counts, expected proportions, and degrees of freedom 289

Chapter 9:

contingency.expected()

Calculates expected values for a contingency table provided as an array or data frame 289

oddsratio()

Calculates an odds ratio and its confidence interval given the numbers of 'successes' and 'failures' in two samples 290

Chapter 10:

p.outside()

Calculates the area of a normal curve outside the interval upper - lower 290

Chapter 11:

CI.var()

Calculates sample variances & standard deviation, and a confidence interval for each, from a numeric vector 291

levene.test()

Executes a Levene test for homogeneity of variances given a numerical vector and a grouping variable of the same length 291

onesample.t()

Executes a 1-sample *t*-test from previously-calculated descriptive statistics: arguments are a hypothesized mean, a sample mean, a sample standard deviation, and a sample size 292

Chapter 12:

twosample.t()

Executes a 2-sample *t*-test (assuming homogeneous variances) from previously-calculated descriptive statistics: arguments the mean, standard deviation, and sample size from two samples 292

Chapter 13:

sign.test ()

Executes a sign test to test whether the median of *x* could equal some specified value 293

Chapter 16:

CI.r()

Calculates confidence limits for a previously-calculated Pearson correlation coefficient, given values for *r* and the sample size. Uses the Fisher approximation 294

Preface

This manual is designed to teach people to use the statistical software S-Plus and to support the process of learning statistical concepts and methods. It is most useful as a workbook to accompany Whitlock and Schluter's *The Analysis of Biological Data*, published by Roberts & Company, Colorado. Although we include enough statistical background to put the procedures being demonstrated in context, we assume that readers will be acquiring most of their understanding of statistical concepts elsewhere.

Several of the authors of this manual have been teaching introductory biostatistics to undergraduate and postgraduate students on two campuses in Australia for more than a decade (in fact one of us, who would prefer not to be identified, taught a biostatistics course for the first time more than three decades ago). In 2008 we discovered the textbook *The Analysis of Biological Data* (referred to in this manual as ABD). We liked everything about the book: its explanations were beautifully clear and aimed at students much like our own; it used a wide variety of real biological examples; it emphasized concepts and procedures important to biologists and explained how they worked; and it introduced some newer computer-intensive techniques that almost all beginning researchers find themselves needing sooner rather than later. We immediately adopted the book as a text for our own introductory biostatistics course. But this adoption acted as a trigger for making some other changes to our teaching—and in particular, to the way we introduced students to statistical software.

To statistical novices, no statistical software is 'user-friendly', and its use needs to be introduced in a structured way which runs in parallel with their acquisition of statistical understanding. At the same time, teaching effort needs to stay focused on statistics rather than software, so that students do not come to see learning to use the software as their primary goal. This manual is intended to allow users to learn to use the software on their own, while keeping a focus on the concepts and procedures which it supports.

We have followed the ABD approach and layout very closely—indeed, we started out with the intention of simply demonstrating in S-Plus every example used in the body of that text. In the end, because everyone has a slightly different view of what should be included in a first statistics course, we added a number of other examples, mostly using our own data, to demonstrate software capabilities that would not otherwise have been covered.

Why S-Plus?

There are a lot of statistical software options, and most of them will execute all the procedures needed in an introductory course. In choosing a software package, we had four criteria beyond its ability to execute procedures taught in the course:

- **It should have little or no cost to students, and should run on operating systems that students are likely to use on their own machines.** Some of us (OK, one of us) remembered teaching statistics in the days when the only computing aid available to students was a hand calculator (the rest of us at least remember being taught that way). While we did not wish to return to those days, they had one huge advantage—students could work on the material anywhere and any time—not just in computer laboratories provided by the university.

Many of our students are part-time, and some are in remote locations. While we can now reasonably expect that students will have access to a computer at home, we cannot reasonably expect them to buy expensive software for themselves. That meant that if we wanted students to work off-campus, we needed to choose software which was either free or very cheap, or which gave students access on their own machines as part of the university's site licence.

- **It should be useful beyond the course.** We wanted students to use professional-quality software that they would not 'grow out of': providing access to all or most of the techniques they were likely to use throughout their careers; and able to import and export data in a wide range of formats (including text files, databases, spreadsheets, and other statistical software).
- **It should have a very strong graphics capability.** We wanted students to realize as quickly as possible that nothing substitutes for an intimate familiarity with the data they are analysing—and easily-usable graphics allow the data to be explored more quickly and thoroughly than anything else. We wanted the graphics capability to cover the whole range from quick-and-dirty exploratory plots to presentation and publication-quality graphs.
- **It should reinforce the statistical concepts we wanted students to grasp, and not get in the way of learning them.** We wanted to avoid both excessive or inappropriate output, and too much difficulty in using the software itself. Excessive output is often a problem with menu-driven software, which may be relatively easy to use¹, but often provides pages of output that users neither asked for nor know what to do with. Especially for novices, our preference was for software that gives users exactly what they request and offers warnings (or refuses to perform) when what they request is questionable. We believe that someone learning to use statistical procedures should also learn to think about what they are doing and work out exactly what it is they want, rather than making guesses about what button to click in the hope that something useful will happen. On the other hand, if software is too difficult to use, students will inevitably concentrate on learning the mechanics of how to use it rather than developing more fundamental understanding.

In the end we chose S-Plus as the best fit to our needs. That choice committed us to producing this manual: there are some excellent introductory books available for S-Plus, but none that we investigated is targeted at undergraduates who begin as complete statistical novices. S-Plus is very powerful and flexible, has superb editable graphics, and its site licence for universities gave enrolled students permission to use the software on their own computers. It had the additional advantage that it provides both a professional-quality graphical user interface (GUI), and an easily-accessible command language which is very similar to that used by the free open-source software R. Mostly we use the GUI, but we also decided to provide a parallel introduction to the command line and to writing basic scripts. Because the command language of S-Plus and R is so similar, students who use this manual should be able to move fairly painlessly to R by the end of it, if they need to do so.

¹ *Ease of use* is relative. We previously used software whose main selling feature was ease of use. But unless we spent considerable time teaching students to use it, they did not cope well. In our experience, the use of any statistical software needs to be introduced to beginners in a structured way in parallel with their statistical understanding.

How to use the manual

If you are a student using ABD as a text, and you have access to S-Plus, you can use S-Plus to work through each chapter of the manual independently. Every example is demonstrated in enough detail for you to carry it out on your own after reading the ABD chapter and/or covering the statistical concepts in class. You should execute every example yourself to make sure that you can carry out the procedures correctly and get to the right result. A set of exercises is provided at the end of each chapter for you to test your skills. You should make sure that you can do them all—you may require some assistance from your instructor to complete some of them successfully. All the data and scripts required for each chapter are available on the CD provided with the manual.

The first chapter of the manual is a basic introduction to S-Plus, and is one of a few chapters whose content is not linked to ABD. The second chapter introduces you to S-Plus graphics. The remaining chapters can be covered in several different orders, but you need to work through these two first. Not all the material in later chapters will necessarily be included in an introductory course.

In most chapters, we show how to execute statistical procedures using both the GUI and the command line. A few procedures require the use of scripts (short programs written in the S language). Where this is the case, we provide the scripts on the CD—and we show you how to load and use them (but we also encourage you to learn to write your own). In many cases, there are more efficient or elegant ways to write scripts than we have used here—in general, we have tried to produce scripts whose logic can be easily understood by beginners, rather than trying for maximum computational efficiency. Appendix 1 provides a more extensive summary of the S language and S-Plus functions relevant to each chapter of the manual.

To the instructor

We believe that learning statistics is like learning to play the piano—there is no substitute for practice. Consequently, in our own teaching, we provide a lot of incentives for students to practice.

In the introductory course that we teach, we expect students to have worked through the appropriate chapter(s) in the manual and attempted the exercises *before* they arrive at the relevant practical class or tutorial—and the first 20 minutes of each 2-hour practical class includes a simple open-book practical test, marked in class, which requires them to analyse some new data using techniques covered in the chapter. (By the end of the course, most students score full marks on most of these tests.) We also run formal (but also open-book) practical exams twice during the course, where the emphasis is on demonstrating that students can make sensible decisions about what to do as well as demonstrating that they can do it. These are also graded immediately. Because students can take this manual—or anything else—into practical tests and exams, we are explicitly *not* testing how well they remember what buttons to press.

When we first changed to this very assessment-oriented approach to the acquisition of practical skills, one unexpected result was that the average grade on the theory exam at the end of the course (which was in the same format and covered the same material as previously) was significantly higher than that achieved by any previous class). Perhaps the development of practical skills really does improve theoretical understanding.

You will notice a scattering of these shaded boxes throughout the manual. In general they contain material we think you will need around that point, but which is not immediately essential to the procedure being demonstrated. This is a practical manual, so where there are no shaded boxes, there are wide margins where you should not hesitate to add your own notes.

Acknowledgements

As noted above, the structure and content of this manual owes a huge debt to Whitlock and Schluter's text, which provides the best introduction we know of to statistical methods for biology students. We are also very grateful to the students in our 2009 biometrics class, and especially to the practical class tutors (Clwedd Burns, Gavin Coombes, Rie Hagihara, and Philip Newey) whose combined input and feedback improved the manual immensely. Finally, for all his help our thanks to Kris Angelovski of SolutionMetrics Pty Ltd, the Australian distributor of S Plus.

Typesetting conventions used in the manual

To make it easier to use the manual, there are several conventions you should note.

Navigating – The instructions about how to navigate around S-Plus are always given in a particular typeface. For example, the way to navigate from the drop-down menus looks like:

Statistics > Regression > Log-linear (Poisson)...

Similarly, this typeface is used to indicate the parts of a dialogue box you need to change.

Entering new information – Where you are required to enter a name or value in a dialogue box or change an existing name, such as perhaps the name of a column (vector) in a data set, the instruction might look like:

Right click at the top of the `No.deaths` column, select **Properties**, and enter **Number of cases (Frequency)** in the **Description** box.

This typeface represents a name or value that you can either enter or change, whereas **the navigation typeface** cannot be altered.

Coding – You learn in Chapter 1 the significance of single lines of code. When programming, pressing the **Enter** key always means something. Where we have reproduced lines of code that you will see on your computer screen, it has often been necessary to spread them over more than one line because the printed page is narrower than your screen. Where this has happened, the subsequent lines of that instruction have been indented. The following example shows two separate instructions (in a lighter shade) and the screen output.

```
titanic.array = table(titanicAdults$Gender,
                     titanicsAdults$Outcome)
titanic.array
      Died Survived
Female  109      316
Male   1329      338
```

Quotation marks when used in code should be straight (`"`) and not curly (`"`), sometimes called “smart” quotes. In S-Plus code, curly quotes will generate an error message.

These conventions will become clear as you work through the chapters.