Statistics

Statistics

An Introduction using R

Michael J. Crawley
Imperial College London, UK

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Crawley, Michael J.
Statistics : an introduction using R / M. J. Crawley.
p. cm.
ISBN 0-470-02297-3 (acid-free : hardback) – ISBN 0-470-2298-1 (acid-free : pbk.)
1. Mathematical statistics–Textbooks. 2. R (Computer program language)
I. Title.
QA276.12.C73 2005
519.5–dc22 2004026793

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-02297-3 (Cloth)
ISBN 0-470-02298-1 (Paper)

Typeset in 10/12pt Times by Thomson Press (India) Limited, New Delhi, India.
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface *xi*

Chapter 1 Fundamentals **1**

Everything Varies 2
Significance 3
Good and Bad Hypotheses 3
Null Hypotheses 3
p Values 3
Interpretation 4
Statistical Modelling 4
Maximum Likelihood 5
Experimental Design 7
The Principle of Parsimony (Occam's Razor) 7
Observation, Theory and Experiment 8
Controls 8
Replication: It's the *n*'s that Justify the Means 8
How Many Replicates? 9
Power 9
Randomization 10
Strong Inference 12
Weak Inference 12
How Long to Go On? 13
Pseudoreplication 13
Initial Conditions 14
Orthogonal Designs and Non-orthogonal Observational Data 14

Chapter 2 Dataframes **15**

Selecting Parts of a Dataframe: Subscripts 19
Sorting 20
Saving Your Work 22
Tidying Up 22

Chapter 3 Central Tendency 23

Getting Help in R 31

Chapter 4 Variance 33

Degrees of Freedom 36
Variance 37
A Worked Example 39
Variance and Sample Size 42
Using Variance 43
A Measure of Unreliability 44
Confidence Intervals 45
Bootstrap 46

Chapter 5 Single Samples 51

Data Summary in the One Sample Case 51
The Normal Distribution 55
Calculations using z of the Normal Distribution 60
Plots for Testing Normality of Single Samples 64
Inference in the One-sample Case 65
Bootstrap in Hypothesis Testing with Single Samples 66
Student's t-distribution 67
Higher-order Moments of a Distribution 69
Skew 69
Kurtosis 71

Chapter 6 Two Samples 73

Comparing Two Variances 73
Comparing Two Means 75
Student's t-test 76
Wilcoxon Rank Sum Test 79
Tests on Paired Samples 81
The Sign Test 83
Binomial Tests to Compare Two Proportions 84
Chi-square Contingency Tables 85
Fisher's Exact Test 90
Correlation and Covariance 93
Data Dredging 95
Partial Correlation 96
Correlation and the Variance of Differences Between Variables 97
Scale-dependent Correlations 98
Kolmogorov-Smirnov Test 100

Chapter 7 Statistical Modelling 103

The Steps Involved in Model Simplification 105
Caveats 106
Order of Deletion 106
Model Formulae in R 106
Interactions Between Explanatory Variables 108
Multiple Error Terms 109
The Intercept as Parameter 1 109
Update in Model Simplification 110
Examples of R Model Formulae 110
Model Formulae for Regression 111
GLMs: Generalized Linear Models 113
The Error Structure 114
The Linear Predictor 115
Fitted Values 116
The Link Function 116
Canonical Link Functions 117
Proportion Data and Binomial Errors 117
Count Data and Poisson Errors 118
GAMs: Generalized Additive Models 119
Model Criticism 119
Summary of Statistical Models in R 120
Model Checking 121
Non-constant Variance: Heteroscedasticity 122
Non-Normality of Errors 122
Influence 123
Leverage 123
Mis-specified Model 124

Chapter 8 Regression 125

Linear Regression 128
Linear Regression in R 129
Error Variance in Regression: $SSY = SSR + SSE$ 136
Measuring the Degree of Fit, r^2 142
Model Checking 143
Polynomial Regression 145
Non-linear Regression 149
Testing for Humped Relationships 152
Generalized Additive Models (gams) 152

Chapter 9 Analysis of Variance 155

One-way Anova 155
Shortcut Formula 161

Effect Sizes 163
Plots for Interpreting One-way Anova 167
Factorial Experiments 171
Pseudoreplication: Nested Designs and Split Plots 175
Split-plot Experiments 176
Random Effects and Nested Designs 178
Fixed or Random Effects? 179
Removing the Pseudoreplication 180
Analysis of Longitudinal Data 180
Derived Variable Analysis 181
Variance Components Analysis (VCA) 181
What is the Difference Between Split-plot and Hierarchical Samples? 185

Chapter 10 Analysis of Covariance 187

Chapter 11 Multiple Regression 195

A Simple Example 195
A More Complex Example 202
Automating the Process of Model Simplification Using step 208
AIC (Akaike's Information Criterion) 208

Chapter 12 Contrasts 209

Contrast Coefficients 210
An Example of Contrasts in R 211
A Priori Contrasts 212
Model Simplification by Step-wise Deletion 214
Contrast Sums of Squares by Hand 217
Comparison of the Three Kinds of Contrasts 218
Aliasing 222
Contrasts and the Parameters of Ancova Models 223
Multiple Comparisons 226

Chapter 13 Count Data 227

A Regression with Poisson Errors 227
Analysis of Deviance with Count Data 229
The Danger of Contingency Tables 234
Analysis of Covariance with Count Data 237
Frequency Distributions 240

Chapter 14 Proportion Data 247

Analyses of Data on One and Two Proportions 249
Count Data on Proportions 249

Odds 250
Overdispersion and Hypothesis Testing 251
Applications 253
Logistic Regression with Binomial Errors 253
Proportion Data with Categorical Explanatory Variables 255
Analysis of Covariance with Binomial Data 260

Chapter 15 Death and Failure Data 263

Survival Analysis with Censoring 265

Chapter 16 Binary Response Variable 269

Incidence Functions 271
Ancova with a Binary Response Variable 275

Appendix 1: Fundamentals of the R Language 281

R as a Calculator 281
Assigning Values to Variables 282
Generating Repeats 283
Generating Factor Levels 283
Changing the Look of Graphics 284
Reading Data from a File 286
Vector Functions in R 287
Subscripts: Obtaining Parts of Vectors 288
Subscripts as Logical Variables 289
Subscripts with Arrays 289
Subscripts with Lists 291
Writing Functions in R 292
Sorting and Ordering 292
Counting Elements within Arrays 294
Tables of Summary Statistics 294
Converting Continuous Variables into Categorical Variables Using cut 295
The split Function 295
Trellis Plots 297
The xyplot Function 299
Three-dimensional (3-D) Plots 300
Matrix Arithmetic 301
Solving Systems of Linear Equations 304

References and Further Reading *305*

Index *309*

Preface

This book is an introduction to the essentials of statistical analysis for students who have little or no background in mathematics or statistics. The audience includes first or second year undergraduate students in science, engineering, medicine and economics, along with post-experience and other mature students who want to re-learn their statistics, or to switch to the powerful new language of R.

For many students, statistics is the least favourite course of their entire time at university. Part of this is because some students have convinced themselves that they are no good at sums, and consequently have tried to avoid contact with anything remotely quantitative in their choice of subjects. They are dismayed, therefore, when they discover that the statistics course is compulsory. Another part of the problem is that statistics is often taught by people who have absolutely no idea how difficult some of the material is for non-statisticians. As often as not, this leads to a recipe-following approach to analysis, rather than to any attempt to understand the issues involved and how to deal with them.

The approach adopted here involves virtually no statistical theory. Instead, the assumptions of the various statistical models are discussed at length, and the practice of exposing statistical models to rigorous criticism is encouraged. A philosophy of model simplification is developed in which the emphasis is placed on estimating effect sizes from data, and establishing confidence intervals for these estimates. The role of hypothesis testing at an arbitrary threshold of significance like $\alpha = 0.05$ is played down. The text starts from absolute basics and assumes absolutely no background in statistics or mathematics.

As to presentation, the idea is that background material would be covered in a series of 1 hour lectures, then this book could be used as a guide to the practical sessions and for homework, with the students working on their own at the computer. My experience is that the material can be covered in 10 to 30 lectures, depending on the background of the students and the depth of coverage it is hoped to achieve. The practical work is designed to be covered in 10 to 15 sessions of about 1.5 hours each, again depending on the ambition and depth of the coverage, and on the amount of one-to-one help available to the students as they work at their computers.

R and S-PLUS

The R language of statistical computing has an interesting history. It evolved from the S language, which was first developed at AT&T's Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. Their idea was to provide a software tool for professional

statisticians who wanted to combine state-of-the-art graphics with powerful model-fitting capability. S is made up of three components. First and foremost, it is a powerful tool for statistical modelling. It enables you to specify and fit statistical models to your data, assess the goodness of fit and display the estimates, standard errors and predicted values derived from the model. It provides you with the means to define and manipulate your data, but the way you go about the job of modelling is not predetermined, and the user is left with maximum control over the model-fitting process. Second, S can be used for data exploration, in tabulating and sorting data, in drawing scatter plots to look for trends in your data, or to check visually for the presence of outliers. Third, it can be used as a sophisticated calculator to evaluate complex arithmetic expressions, and a very flexible and general object-oriented programming language to perform more extensive data manipulation. One of its great strengths is in the way in which it deals with vectors (lists of numbers). These may be combined in general expressions, involving arithmetic, relational and transformational operators such as sums, greater-than tests, logarithms or probability integrals. The ability to combine frequently-used sequences of commands into functions makes S a powerful programming language, ideally suited for tailoring one's specific statistical requirements. S is especially useful in handling difficult or unusual data sets, because its flexibility enables it to cope with such problems as unequal replication, missing values, non-orthogonal designs, and so on. Furthermore, the open-ended style of S is particularly appropriate for following through original ideas and developing new concepts. One of the great advantages of learning S is that the simple concepts that underlie it provide a unified framework for learning about statistical ideas in general. By viewing particular models in a general context, S highlights the fundamental similarities between statistical techniques and helps play down their superficial differences. As a commercial product S evolved into S-PLUS, but the problem was that S-PLUS was very expensive. In particular, it was much too expensive to be licensed for use in universities for teaching large numbers of students. In response to this, two New Zealand-based statisticians, Ross Ihaka and Robert Gentleman from the University of Auckland, decided to write a stripped-down version of S for teaching purposes. The letter R 'comes before S' so what would be more natural than for two authors whose first initial was 'R' to christen their creation R. The code for R was released in 1995 under a GPL (General Public License), and the core team was rapidly expanded to 15 members (they are listed on the web site, below). Version 1.0.0 was released on 29 February 2000. This book is written using version 1.8.1, but all the code will run under R 2.0.0 (released in September 2004). R is an Open Source implementation of S-PLUS, and as such can be freely downloaded. If you type CRAN into your Google window you will find the site nearest to you from which to download it. Or you can go directly to

http://cran.r-project.org

There is a vast network of R users world-wide, exchanging functions with one another, and a vast resource of libraries containing data and programs. There is a useful journal called *R News* that you can read at CRAN.

This book has its own web site at

http://www.imperial.ac.uk/bio/research/crawley/statistics

Here you will find all the data files used in the text; you can download these to your hard disc and then run all of the examples described in the text. The executable statements are shown in the text in Arial font. There are files containing all the commands for each chapter, so you can paste the code directly into R instead of typing it from the book. Another file supplies the code necessary to generate all of the book's figures. There is a series of 14 fully-worked stand-alone practical sessions covering a wide range of statistical analyses. Learning R is not easy, but you will not regret investing the effort to master the basics.

M. J. Crawley
Ascot