Statistics

An Introduction using R

Michael J. Crawley
Imperial College London, UK

John Wiley & Sons, Ltd

Copyright © 2005 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

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

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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 0-470-02298-1 (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)

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

1

Fundamentals

The hardest part of any statistical work is getting started – and one of the hardest things about getting started is choosing the right kind of statistical analysis. The choice depends on the nature of your data and on the particular question you are trying to answer. The truth is that there is no substitute for experience; the way to know what to do, is to have done it properly lots of times before.

The key is to understand what kind of **response** variable you have got, and to know the nature of your **explanatory** variables. The response variable is the thing you are working on; it is the variable whose variation you are attempting to understand. This is the variable that goes on the y axis of the graph (the ordinate). The explanatory variable goes on the x axis of the graph (the abscissa); you are interested in the extent to which variation in the response variable is associated with variation in the explanatory variable. A continuous measurement is a variable like height or weight that can take any real numbered value. A categorical variable is a **factor** with two or more **levels**: gender is a factor with two levels (male and female), and a rainbow might be a factor with seven levels (red, orange, yellow, green, blue, indigo, violet).

It is essential, therefore, that you know:

- which of your variables is the response variable;
- which are the explanatory variables;
- are the explanatory variables continuous or categorical, or a mixture of both;
- what kind of response variable have you got – is it a continuous measurement, a count, a proportion, a time-at-death or a category?

These simple keys will then lead you to the appropriate statistical method.

1. The explanatory variables

(a)	All explanatory variables continuous	**Regression**
(b)	All explanatory variables categorical	**Analysis of variance (Anova)**
(c)	Explanatory variables both continuous and categorical	**Analysis of covariance (Ancova)**

Statistics: An Introduction using R M. J. Crawley
© 2005 John Wiley & Sons, Ltd ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

2. *The response variable*

(a)	Continuous	**Normal regression, Anova or Ancova**
(b)	Proportion	**Logistic regression**
(c)	Count	**Log linear models**
(d)	Binary	**Binary logistic analysis**
(e)	Time-at-death	**Survival analysis**

There are some key ideas that need to be understood from the outset. We cover these here before getting into any detail about different kinds of statistical model.

Everything Varies

If you measure the same thing twice you will get two different answers. If you measure the same thing on different occasions you will get different answers because the thing will have aged. If you measure different individuals, they will differ for both genetic and environmental reasons (nature and nurture). Heterogeneity is universal: spatial heterogeneity means that places always differ and temporal heterogeneity means that times always differ.

Because everything varies, finding that things vary is simply not interesting. We need a way of discriminating between variation that is scientifically interesting, and variation that just reflects background heterogeneity. That is why we need statistics. It is what this whole book is about.

The key concept is the amount of variation that we would expect to occur by chance alone, when nothing scientifically interesting was going on. If we measure bigger differences than we would expect by chance, we say that the result is statistically significant. If we measure no more variation than we might reasonably expect to occur by chance alone, then we say that our result is not statistically significant. It is important to understand that this is not to say that the result is not important. Non-significant differences in human life span between two drug treatments may be massively important (especially if you are the patient involved). Non-significance is not the same as 'not different'. The lack of significance may simply be due to the fact that our replication is too low.

On the other hand, when nothing really **is** going on, then we want to know this. It makes life much simpler if we can be reasonably sure that there is no relationship between y and x. Some students think that 'the only good result is a significant result'. They feel that their study has somehow failed if it shows that 'A has no significant effect on B'. This is an understandable failing of human nature, but it is not good science. The point is that we want to know the truth, one way or the other. We should try not to care too much about the way things turn out. This is not an amoral stance, it just happens to be the way that science works best. Of course, it is hopelessly idealistic to pretend that this is the way that scientists really behave. Scientists often hope passionately that a particular experimental result will turn out to be statistically significant, so that they can have a paper published in Nature and get promoted, but that doesn't make it right.

Significance

What do we mean when we say that a result is significant? The normal dictionary definitions of significant are 'having or conveying a meaning' or 'expressive; suggesting or implying deeper or unstated meaning' but in statistics we mean something very specific indeed. We mean that 'a result was unlikely to have occurred by chance'. In particular, we mean 'unlikely to have occurred by chance if the null hypothesis was true'. So there are two elements to it: we need to be clear about what we mean by 'unlikely', and also what exactly we mean by the 'null hypothesis'. Statisticians have an agreed convention about what constitutes 'unlikely'. They say that an event is unlikely if it occurs less than 5% of the time. In general, the 'null hypothesis' says that 'nothing's happening' and the alternative says 'something **is** happening'.

Good and Bad Hypotheses

Karl Popper was the first to point out that a good hypothesis is one that is capable of **rejection**. He argued that **a good hypothesis is a falsifiable hypothesis**. Consider the following two assertions.

1. There are vultures in the local park.
2. There are no vultures in the local park.

Both involve the same essential idea, but one is refutable and the other is not. Ask yourself how you would refute option 1. You go out into the park and you look for vultures, but you don't see any. Of course, this doesn't mean that there aren't any. They could have seen you coming, and hidden behind you. No matter how long or how hard you look, you cannot refute the hypothesis. All you can say is 'I went out and I didn't see any vultures'. One of the most important scientific notions is that **absence of evidence is not evidence of absence**. Option 2 is fundamentally different. You reject hypothesis 2 the first time that you see a vulture in the park. Until the time that you **do** see your first vulture in the park, you work on the assumption that the hypothesis is true. But if you see a vulture, the hypothesis is clearly false, so you reject it.

Null Hypotheses

The null hypothesis says 'nothing's happening'. For instance, when we are comparing two sample means, the null hypothesis is that the means of the two samples are the same. Again, when working with a graph of y against x in a regression study, the null hypothesis is that the slope of the relationship is zero, i.e. y is not a function of x, or y is independent of x. The essential point is that the null hypothesis is falsifiable. We reject the null hypothesis when our data show that the null hypothesis is sufficiently unlikely.

p Values

A p value is an estimate of the probability that a particular result, or a result more extreme than the result observed, could have occurred by chance, if the null hypothesis were true. In short, the p value is a measure of the credibility of the null hypothesis. If

something is very unlikely to have occurred by chance, we say that it is statistically significant, e.g. $p < 0.001$. For example, in comparing two sample means, where the null hypothesis is that the means are the same, a low p value means that the hypothesis is unlikely to be true and the difference is statistically significant. A large p value (e.g. $p = 0.23$) means that there is no compelling evidence on which to reject the null hypothesis. Of course, saying 'we do not reject the null hypothesis' and 'the null hypothesis is true' are two quite different things. For instance, we may have failed to reject a false null hypothesis because our sample size was too low, or because our measurement error was too large. Thus, p values are interesting, but they don't tell the whole story; effect sizes and sample sizes are equally important in drawing conclusions.

Interpretation

It should be clear by this point that we can make two kinds of mistakes in the interpretation of our statistical models:

- we can reject the null hypothesis when it is true, or
- we can accept the null hypothesis when it is false.

These are referred to as **Type I** and **Type II** errors respectively. Supposing we knew the true state of affairs (which, of course, we seldom do), then in tabular form:

	Actual situation	
Null hypothesis	*True*	*False*
Accept	Correct decision	Type II
Reject	Type I	Correct decision

Statistical Modelling

The object is to determine the values of the parameters in a specific model that lead to the best fit of the model to the data. The data are sacrosanct, and they tell us what actually happened under a given set of circumstances. It is a common mistake to say 'the data were fitted to the model' as if the data were something flexible, and we had a clear picture of the structure of the model. On the contrary, what we are looking for is the minimal adequate model to describe the data. The model is fitted to the data, not the other way around. The best model is the model that produces the least unexplained variation (the **minimal residual deviance**), subject to the constraint that all the parameters in the model should be statistically significant.

You have to specify the model. It embodies your mechanistic understanding of the factors involved, and of the way that they are related to the response variable. We want the model to be **minimal** because of the principle of parsimony, and **adequate** because there is no point in retaining an inadequate model that does not describe a significant fraction of the variation in the data. It is very important to understand that there is not just

one model; this is one of the common implicit errors involved in traditional regression and Anova, where the same models are used, often uncritically, over and over again. In most circumstances, there will be a large number of different, more or less plausible models that might be fitted to any given set of data. Part of the job of data analysis is to determine which, if any, of the possible models are adequate and then, out of the set of adequate models, which is the minimal adequate model. In some cases there may be no single best model and a set of different models may all describe the data equally well (or equally poorly if the variability is great).

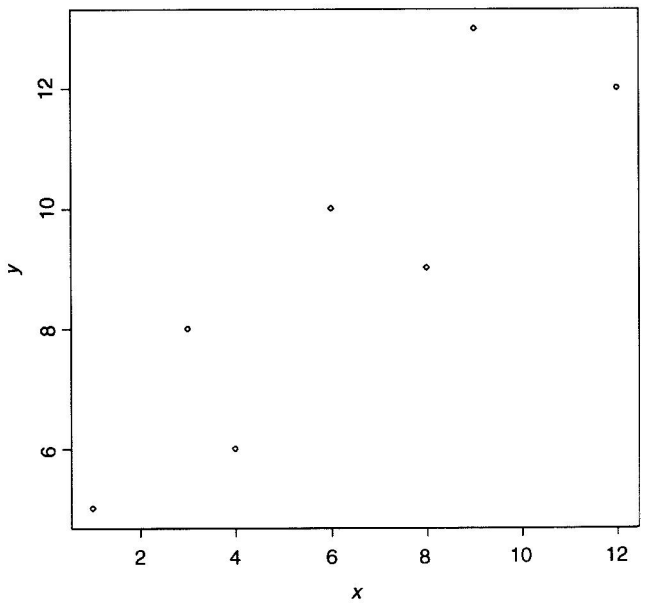
Maximum Likelihood

What exactly do we mean when we say that the parameter values should afford the 'best fit of the model to the data'? The convention we adopt is that our techniques should lead to **unbiased, variance minimizing estimators**. We define 'best' in terms of **maximum likelihood**. This notion is likely to be unfamiliar, so it is worth investing some time to get a feel for it. This is how it works.

- Given the data,
- and given our choice of model,
- what values of the parameters of that model make the observed data most likely?

Here are the data: y is the response variable and x is the explanatory variable. Because both x and y are continuous variables, the appropriate model is regression.

```
x<-c(1,3,4,6,8,9,12)
y<-c(5,8,6,10,9,13,12)
plot(x,y)
```

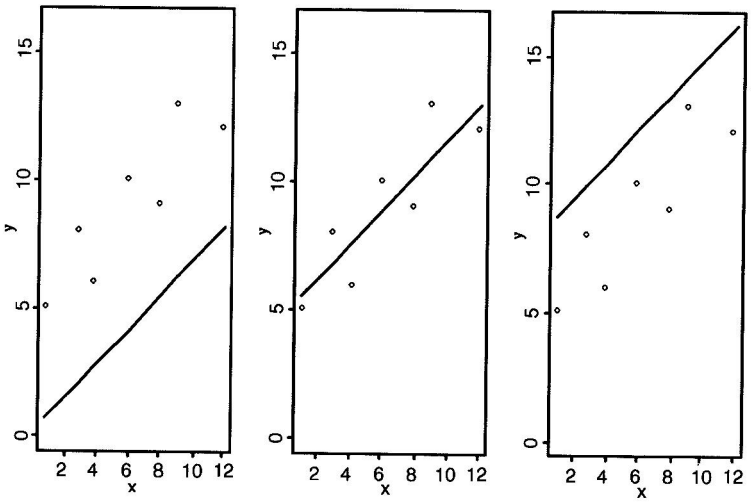


Now we need to select a regression model to describe these data from the vast range of possible models available. Let's choose the simplest model, the straight line

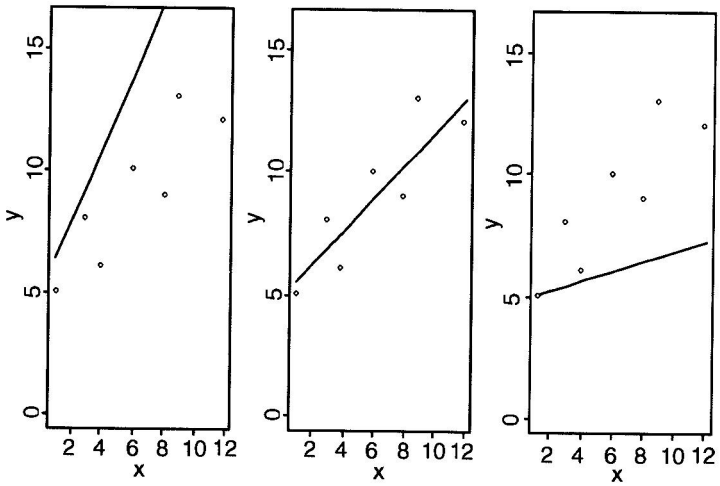
$$y = a + bx.$$

This is a two-parameter model; the first parameter, a, is the intercept (the value of y when x is 0) and the second, b, is the slope (the change in y associated with unit change in x). The response variable y, is a linear function of the explanatory variable x. Now suppose that we knew that the slope was 0.68, then the maximum likelihood question can be applied to the intercept a.

If the intercept were 0 (left-hand graph, below), would the data be likely? The answer of course, is no. If the intercept were 8 (right-hand graph) would the data be likely? Again, the answer is obviously no. The maximum likelihood estimate of the intercept is shown in the central graph (its value turns out to be 4.827).



We could have a similar debate about the slope. Suppose we knew that the intercept was 4.827, then would the data be likely if the graph had a slope of 1.5 (left graph, below)?



The answer, of course, is no. What about a slope of 0.2 (right graph)? Again, the data are not at all likely if the graph has such a gentle slope. The maximum likelihood of the data given the model is obtained with a slope of 0.679 (centre graph). This is not how the procedure is actually carried out, but it makes the point that we judge the model on the basis of how likely the data would be if the model were correct. In practice of course, both parameters are estimated simultaneously.

Experimental Design

There are only two key concepts:

- replication, and
- randomization.

You replicate to increase reliability. You randomize to reduce bias. If you replicate thoroughly and randomize properly, you will not go far wrong.

There are a number of other issues whose mastery will increase the likelihood that you analyse your data the right way rather than the wrong way:

- the principle of parsimony,
- the power of a statistical test,
- controls,
- spotting pseudoreplication and knowing what to do about it,
- the difference between experimental and observational data (non-orthogonality).

It does not matter very much if you cannot do your own advanced statistical analysis. If your experiment is properly designed, you will often be able to find somebody to help you with the statistics. However, if your experiment is not properly designed, or not thoroughly randomized, or lacking adequate controls, then no matter how good you are at statistics, some (or possibly even all) of your experimental effort will have been wasted. No amount of high-powered statistical analysis can turn a bad experiment into a good one. R is good, but not that good.

The Principle of Parsimony (Occam's Razor)

One of the most important themes running through this book concerns model simplification. The principle of parsimony is attributed to the 14th century English Nominalist philosopher William of Occam who insisted that, given a set of equally good explanations for a given phenomenon, then **the correct explanation is the simplest explanation**. It is called Occam's razor because he 'shaved' his explanations down to the bare minimum. In statistical modelling, the principle of parsimony means that:

- models should have as few parameters as possible,
- linear models should be preferred to non-linear models,
- experiments relying on few assumptions should be preferred to those relying on many,