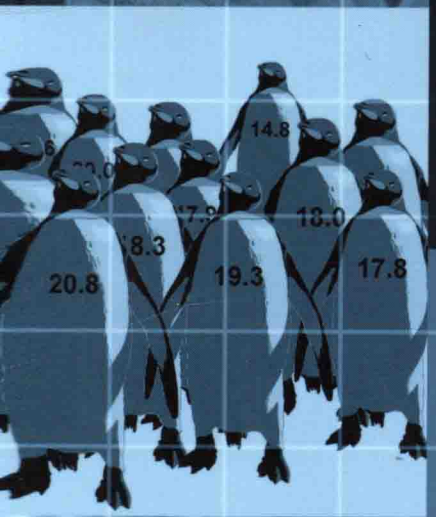
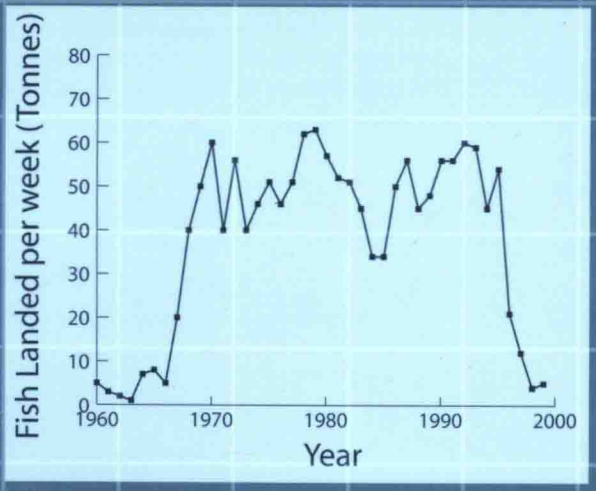


Introduction to Statistics for Biology

THIRD EDITION

Robin H. McCleery
Trudy A. Watt
Tom Hart



Chapman & Hall/CRC
Taylor & Francis Group

Introduction to Statistics for Biology

THIRD EDITION

Robin H. McCleery
Trudy A. Watt
Tom Hart

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-652-8 (Softcover)
International Standard Book Number-13: 978-1-58488-652-5 (Softcover)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Colyear Dawkins

Contributors

Dr. Robin McCleery, Edward Grey Institute, Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, Email: robin.mccleery@zoo.ox.ac.uk, Tel 01865 271161, Fax 01865 271168.

Dr. Trudy A. Watt, Trinity College, Broad Street, Oxford, OX1 3BH, Email: trudy.watt@trinity.ox.ac.uk, Tel 01865 279881, Fax 01865 279911.

Mr. Tom Hart, 1. Imperial College at Silwood Park, Ascot, Berks, SL5 7PY, Email: tom.hart@imperial.ac.uk, Tel. 0207 594 2447, (no fax); 2. British Antarctic Survey, High Cross, Madingley Road, Cambridge, CB3 0ET, Tel: 0207 594 2307.

Preface

Students of biological subjects in further and higher education often lack confidence in their numerical abilities. Those coming straight from school are likely to have some science subjects in their school leaving qualifications but may have little mathematics. Mature students often describe themselves as "rusty," and some biologists choose biological sciences specifically to escape from numbers. Unfortunately for them, an understanding of experimental design and statistics are as central to modern biology as an understanding of evolution. As well as demonstrating the importance of statistics, this text aims to calm fears and to provide a relatively painless way into the subject.

We stress a reliance on computers throughout. Although we demonstrate procedures and concepts in all tests, no one in "the real world" would conceivably do anything but the simplest statistics by hand, or rely solely on a calculator. Modern statistical programs are very good at analysing large data sets and are becoming increasingly good at reporting errors to show when a test might not be appropriate.

The title of the third edition has been shortened from *Introductory Statistics for Biology Students* to the more inclusive *Introductory Statistics for Biology*, which reflects the fact that many researchers refer back to introductory statistics texts to refresh their memories.

The second edition was published in 1997, and by 2007, a new edition was overdue. MINITAB has reached its 15th edition with the incorporation of numerous improvements, and the third edition of this book is set to coincide with its release. These changes are reflected throughout the book both in the range of tests that can be performed and increased clarity to the beginner. Robin McCleery and Tom Hart have joined Trudy Watt as coauthors.

Trudy Watt wrote the first and second editions while a senior lecturer in Statistics and Ecology at Wye College, University of London. Robin McCleery's experience of teaching first-year biology students at Oxford University with the 2nd edition as a textbook has suggested some areas for expansion, some for deletion, and a change in the balance of topics. Tom Hart was a tutor and demonstrator on this course in the Zoology Department, and both have tried to address issues that current students struggle with.

Although the ethos remains the same, there are some significant changes:

- Removal of many of the exercises, which have been replaced by worked examples.
- A new general template for carrying out statistical tests from hypothesis to interpretation. We repeat this throughout to show the generality to different types of tests.

- An emphasis on experimental design, and simulating data prior to carrying out an experiment.
- MINITAB analyses and graphics have been updated to Releases 14 and 15.

In particular, the changes are:

Chapters 1 to 3 have been greatly expanded to provide a more thorough grounding in the basic ideas behind statistical thinking. We explain probability in some detail to clarify the rationale behind hypothesis testing before moving on to simple tests. We have emphasised the common thread running through much of statistics by formulating a general approach to carrying out a statistical test. Some more detailed explanation in the early chapters ensures that we can follow the same template throughout the book.

Chapter 4 represents the chapters on sampling and experimental design from the 2nd edition, which we have combined for simplicity and because they are so intrinsically linked.

Chapters 5 to 7 on Analysis of Variance (ANOVA) are very similar in scope to previous editions but have been rewritten with more emphasis on factorial designs and interactions. We have also removed many references to post-hoc testing to discourage people from excessive use of these tests.

Chapter 8 (Correlation and Regression) has been rewritten to bring it more into line with the approach taken in Chapters 5 to 7, where we initially introduce the method with a very simple numerical example. We have also made the similarities between ANOVA and regression more apparent.

The chapter concerning data from an observational study has been deleted to allow for a more thorough discussion of categorical data and nonparametric statistics in Chapter 9 and Chapter 10. What was previously discussed in the chapter on observational studies has now been partially covered in Chapters 9 to 11, and partly covered by references.

Chapter 11 is now a general project template with advice on how to carry out and write up an undergraduate project. We also include a short sample report to illustrate many of the points we make.

We have included a trial copy of MINITAB version 15 for you to try. Just insert the disk and follow the instructions to get a 30 day free trial. Details of purchase can be found at http://www.minitab.com.

MINITAB has a large number of data sets available for practice use. Descriptions are found by consulting the index, selecting data sets, and clicking on the name of one. To upload a data set into the worksheet, click on "file, open worksheet" and select the desired filename. The reader is encouraged to use these data sets, which replace the end-of-chapter examples in the 2nd edition. For example:

> POTATO.MTW. In this experiment a rot-causing bacterium was injected into potatoes in low, medium, or high amounts (C1). The potatoes were

left for 5 days at 10 or 16 degrees C (C2) and with 2, 6, or 10% oxygen (C3). The diameter of the rotted area on each potato was measured as the response variable (C4).

YIELDSTDV.MTW. In this study, there are eight blocks (C4) and three factors: reaction time (C5), reaction temperature (C6), and catalyst (C7). The yield of the chemical reaction was recorded (C8).

Our inspirations and interest in statistics are varied, but this book remains dedicated to the late Colyear Dawkins whose enthusiasm for communicating the principles of experimental design and analysis was infectious. We would also like to thank a number of people for comments while writing this new edition, in particular, Marian Dawkins, Marta Szulkin, Matt Towers and Liz Masden for comments on the text.

Note to Students

Statistics is an essential tool for all life scientists. One of the most important parts of a college or university course in biology and related subjects is learning to understand and critically evaluate research evidence. Open any scientific journal in the life sciences and you will find the pages littered with probability statements, test statistics, and other jargon that must be understood if you are to make sense of the claims being made by the researchers. Also, as an apprentice scientist yourself, you will soon start to undertake your own investigations and need the right tools for the correct interpretation of the results.

Unfortunately, a first-year statistics course is often seen as just an inconvenient hurdle to be jumped, especially by students who may have little mathematical background. This tends to make people focus on solving problems, which they often do by rote and without perceiving the underlying rationale. In our view this approach actually makes it harder to understand the subject, and we feel that you really need to develop some curiosity about the why's and wherefore's as well as the "how to." One result of this approach is that you do need to read the book from start to finish, rather than dipping in for the bit you want. The argument builds, and we have not always cross referred back to the basic material at every stage.

Many introductory texts contain a disclaimer about the mathematics they will require you to tackle but promptly renege on their promise to shield you from the horrors of equations within a page or two. The fact is that, without some understanding of equations, statistics is going to be hard to explain, but it is also true that "serious mathematics," by which we mean an understanding

of techniques such as calculus, the proof of theorems, and so on, is not really necessary. Most of the ideas involved can be given an intuitive or visual representation, and that is the approach we use in this book.

Another feature of the book is a reliance on the use of computerised methods. There is a tradition of sadism amongst teachers, matched by masochism in some students, who share the idea that unless you have experienced the agony of manual calculation you do not really understand the subject. We dissent from this view for a number of reasons. For some people the effort involved in getting the calculations right gets in the way of understanding. Many manual methods employ algebraic shortcuts to reduce the amount of key pressing on the calculator, but without some facility at algebra, these completely obscure the underlying rationale. It is, of course, true that computers can make things a bit too easy. The numbers go in one end and a printout comes out of the other, but the user may not really have much of a clue what it all means. What we aim to do here is to explain the ideas behind the methods and let the calculating machinery do the drudgery, but you do need to be curious about all the numbers that appear in the output. Most of them are there for a reason, and you should remain uncomfortable until you know what each of them represents.

Although generic computer programs such as spreadsheets contain statistical functions, they often have shortcomings in reliability and in the ease with which statistical information is presented. We have chosen to emphasise the use of specialised statistical packages, and specifically one called MINITAB. Almost everything in this book can be achieved using the student version of this program (Student 14) which is modestly priced.[1] The full release of MINITAB is currently Release 15, and it has a number of facilities not found in the Student release. We have included a time-limited trial copy on CD for you to try. MINITAB has much to recommend it for this purpose, as it was originally written primarily as a teaching tool. It incorporates many useful functions for exploring the ideas behind conventional tests, and contains built-in tutorial, reference, and help materials. Mainstream packages (the full version of MINITAB, SPSS, SAS, Genstat) are expensive for individuals to purchase, though many institutions have site licences that allow registered students to use them at a much lower cost. In practice, there is a high degree of convergence in the way the user interfaces operate, so the skills learned using MINITAB are readily transferred.

Recently, there has been another innovation, the existence of a free statistical package called R (see http://cran.r-project.org). The main problem with this package, in our view, is the user interface, which relies almost entirely on commands typed in from the keyboard. Our experiences teaching with early releases of MINITAB, which worked in a similar way, convince us that most people find the "point and click" interface much easier. However,

[1] MINITAB is a registered trademark of Minitab Inc., whose cooperation we gratefully acknowledge http://www.minitab.com/. Student 14 release details: http://www.minitab.com/products/minitab/student/default.aspx.

MINITAB can help prepare you for something like R because the MINITAB command line lives on behind the scenes (see Appendix A, Section A.10).

Most beginners in statistics find that "choosing the right test" is the hardest bit of the subject and are looking for a "cookbook" or perhaps a taxonomic key to tell them what procedure to follow in a given case. In fact, even experienced statisticians will admit that they sometimes see how to solve a problem by spotting an analogy with an analysis they have encountered elsewhere in a book or journal. We could have written the book in the form of a series of recipes (and we present a taxonomic key in Appendix E) but in our view this would not achieve our objective, which is to encourage you to think statistically. Experienced cooks rarely use a recipe book, except for inspiration, because they have realised that there is only a small number of basic techniques to know. Gravy, béchamel and cheese sauce are variants of a single method; what you need to understand is how to combine a fat, a thickening agent and a liquid without making it lumpy. Similarly, once you start to get the hang of statistical thinking, you realise that there is a single thread running through the whole subject. What at first sight seems to be a book full of disparate tests (many with impressive-sounding names) is a consistent series of methods all based on a common underlying structure of reasoning. Our hope is that, by the time you have finished using this book, you will begin to appreciate this unity in statistical thought and, maybe, just a little, start to share our fascination with it.

Conventions in Our Presentation

Generally, concepts appearing for the first time appear in **bold** and emphasis is made using *italics*. We have also used bold and normal fonts to represent input to MINITAB, as explained in Appendix A.

We have not thought it necessary to number equations, with a few obvious exceptions.

The results of manual calculations are generally given to 2 places of decimals (2 d.p.) rounded off in the usual way. Thus, 2.561 becomes 2.56 (to 2 d.p.), and 2.565 becomes 2.57.

Numbers less than 0.01 are usually given to 3 significant figures. If you see $p = 0.000$ in a computer output, it is unlikely that the probability is really 0! It means $p < 0.0005$.

Some care must be taken when verifying calculations manually using intermediate numbers printed by the computer. The computer often rounds off its output but will do any further calculations using the full numerical values and rounding off the result. This can give rise to apparent rounding errors in some circumstances, but we think we have found all the cases where this might cause confusion.

Contents