Weighing the Odds

Weighing the Odds

A Course in Probability and Statistics

David Williams

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
10 Stamford Road, Oakleigh, VIC 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2001

Printed in the United Kingdom at the University Press, Cambridge

Typeset by the author

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

Library of Congress Cataloguing in Publication data

ISBN 0 521 80356 X hardback
ISBN 0 521 00618 X paperback

For Sheila,
for Jan and Ben, for Mags and Jeff;
and, of course,
for Emma and Sam and awaited 'Bump'
from Gump, the Grumpy Grandpa

(I can *'put those silly sums away' now, Emma.)*

Cambridge University Press
978-0-521-00618-7 - Weighing the Odds: A Course in Probability and Statistics
David Williams
Frontmatter
More information

Contents

Cambridge University Press
978-0-521-00618-7 - Weighing the Odds: A Course in Probability and Statistics
David Williams
Frontmatter
More information

Cambridge University Press
978-0-521-00618-7 - Weighing the Odds: A Course in Probability and Statistics
David Williams
Frontmatter
More information

Preface

Probability and Statistics used to be married; then they separated; then they got divorced; now they hardly ever see each other. (That's a mischievous first sentence, but there is more than an element of truth in it, what with the subjects' having different journals and with Theoretical Probability's sadly being regarded by many – both mathematicians and statisticians – as merely a branch of Analysis.) In part, this book is a move towards much-needed reconciliation. It is written at a difficult time when Statistics is being profoundly changed by the computer revolution, Probability much less so.

This book has many unusual features, as you will see later and as you may suspect already! Above all, it is meant to be fun. I hope that in the Probability you will enjoy the challenge of some 'almost paradoxical' things and the important applications to Genetics, to Bayesian filtering, to Poisson processes, etc. The real-world importance of Statistics is self-evident, and I hope to convey to you that subject's very considerable intrinsic interest, something too often underrated by mathematicians. The challenges of Statistics are somewhat different in kind from those in Probability, but are just as substantial.

(a) *For whom is the book written?* There are different answers to this question. The book derives from courses given at Bath and at Cambridge, and an explanation of the situation at those universities identifies two of the possible (and quite different) types of reader.

At Bath, students first have a year-long gentle introduction to Probability and Statistics; and before they start on the type of Statistical theory presented here, they do a course in Applied Statistics. They are therefore familiar with the mechanics of elementary statistical methods and with some computer packages. Students with this type of background (for whom I provide reminders of some Linear Algebra, etc) will find in this book the unifying ideas which fit the separate Statistical methods into a coherent picture, and also ideas which allow those methods to be extended via modern computer techniques.

At Cambridge, students are introduced to Statistics only when they have a thorough grounding in Mathematics; and they (or at least, many of them) like to see how Linear Algebra, etc, may be applied to real-world problems. I believe it very important to 'sell' Probability and Statistics to students who see themselves as mathematicians; and even to try to 'convert' some of them.

As someone who sees himself as a mathematician (who has always worked in Probability theory and) who very much wishes he knew more Statistics and had greater wisdom in that subject, I hope that I am equipped to sell it to those who would find medians, modes, square-root formulae for standard deviation, and the like, a huge turn-off. (I know that meeting Statistics through these topics came very close to putting me off the subject for life.) You will see that throughout this book I am going to be honest. Mathematics students should know that Probability is just as axiomatic and rigorous as (say) Group Theory, to which it is connected in remarkable ways (see the very end of Chapter 9).

(b) Not only is it the case then that we look at a lot of important topics, but we also look at them seriously – which does not detract from our enjoyment.

So let me convince you just how serious this book is. (The previous sentence was written with a twinkle, of course. But seriously ... .)

I see great value in *both* Frequentist and Bayesian approaches to Statistics; and the last thing I want to do is to dwell to too great an extent on old controversies. However, it is undeniable that there is a profound difference between the two philosophies, even in cases where there is universal agreement about the 'answers'; and, for the sake of clarity, I usually present the two approaches separately. Where there seems to remain controversy (for example in connection with sharp hypotheses in Bayesian theory), I say very clearly what I think. Where I am uneasy about some aspect of the subject, and I *am* about several, I say that too.

In regard to Probability, I explain why the 'definition' of probability in terms of long-term relative frequency is fatally flawed from the point of view of logic (not just impracticality). I also explain very fully why Probability (the subject) only works if we do not attempt to define what probability means in the real world. This gives Mathematics a great advantage over the approach of Philosophy, a seemingly unreasonable advantage since the 'definition' approach seems at first more honest. It is worth quoting Wigner's famous (if rather convoluted) statement:

> The language of mathematics reveals itself [to be] unreasonably effective in the natural sciences ... a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure even though perhaps to our bafflement, to wide branches of learning.

That for the quantum world we use a completely different Quantum Probability calculus will also be explained, the Mathematics of Analysis of Variance (ANOVA) having prepared the ground. It is important to realize that *the* *real world follows which Mathematics it chooses: we cannot insist that Nature follow rules which we think inevitable.*

(c) I said that, as far as Statistics is concerned, this book is being written at a difficult time. Statistics in this new century will be somewhat different in kind from the Statistics of most of last century in that a significant part of the *everyday practice* of Statistics (I am not talking about developments in research) will consist of applying Bayes' formula via MCMC (Monte-Carlo-Markov-Chain) packages, of which `WinBUGS` is a very impressive example. A package `MLwiN` is a more user-friendly package which does quite a lot of classical methods as well as some important MCMC work.

These packages allow us to study more realistic models; and they can deal more-or-less exactly with any sample size in many situations where classical methods can provide only approximate answers for large samples. (However, they can run into serious difficulty, sometimes on very simple problems.)

Many ideas from classical Frequentist Statistics – 'deviance', 'relative entropy', etc – continue to play a key rôle even in 'MCMC' Statistics. Moreover, classical results often serve to cross-check 'modern' ones, and those large-sample results do guarantee the desired consistency if sample sizes were increased. *A broad background in Statistics culture remains as essential as ever.* And the classical Principle of Parsimony ('Always use the **simplest** *acceptable* model') warns us not to be seduced by the availability of remarkable computer packages into using over-'sophisticated' (and, in consequence, possibly non-robust) models with many parameters. Of course, if Science says that our model *requires* many parameters, so be it.

In part, this book is laying foundations on which your Applied Statistics work can build. It does contain a significant amount of numerical work: to illustrate topics and to show how methods and packages work – or fail to work. It discusses some aspects of how to decide if one's model provides an acceptably good 'fit'; and indicates how to test one's model for robustness.

It takes enough of a look to get you started at the computer study of non-classical models. I hope that from such material you will learn useful methods and principles for real applications. The fact that I often 'play devil's advocate', and encourage you to think things out for yourself, should help. But the book is already much longer than originally intended; and because I think that each real example should be taken seriously (and not seen as a trite exercise in the arithmetic of the t-test or ANOVA, for example), I leave the study of real examples to parallel courses of study. *That parallel study of real-world examples should form the main set of 'Exercises' for the Statistics part of this book.* Mathematical exercises are less important; and this book contains very few exercises in which (for example) integration masquerades as Statistics. *In regard to Statistics, common sense and scientific insight remain, as they have always been, more important than Mathematics. But this must never be made an excuse for not using the right Mathematics when it is available.*

(d) The book covers a very limited area. **It is meant only to provide sufficient of a link between Probability and Statistics to enable you to read more advanced books on Statistics written by wiser people, and to show also that 'Probability in its own right'** (the large part of that subject which is 'separate' from Statistics) **is both fascinating and very important.** Sadly, to read more advanced Probability written by wiser people, you will need first to study a lot more Analysis. However, there is much other Probability at this level. See Appendix D.

This *textbook* is meant to *teach*. I therefore always try to provide the most intuitive explanations for theorems, etc, not necessarily the neatest or cleverest ones. The book is *not* a work of scholarship, so that it does not contain extensive references to the literature, though many of the references it mentions *do*, so you can follow up the literature with their help. *Appendix D, 'A small sample of the literature', is one of the key sections in this book. Keep referring to it.* The book is 'modern' in that it recognizes the uses and limitations of computers.

I have tried, as far as possible, to explain *everything* in the limited area covered by the book, including, for example, giving 'C' programs showing how random-number generators work, how statistical tables are calculated, how the MCMC 'Gibbs sampler' operates, etc. I like to know these things myself, so some of you readers might too. But that is not the main reason for giving 'C' details. However good a package is, there will be cases to which it does not apply; one sometimes *needs* to know how one can program things oneself.

In regard to 'C', I think that you will be able to follow the logic of the programs even if you are not familiar with the language. You can easily spot what things are peculiar to 'C' and what really matters. The same applies to `WinBUGS` . I have always believed that *the best way to learn how to program is to read programs*, not books on programming languages! I spare you both pointers and Object-Oriented Programs, so my 'C' in this book is rather basic. There is a brief note in Appendix A about *static* variables which can to some extent be used to obtain some advantages of OOP without the clumsiness.

Apology: I *do* omit explaining why the Gibbs sampler works. Since I have left the treatment of Markov chains to James Norris's fine recent book (which has a brief section on the Gibbs sampler), there is nothing else I can do.

CAMBRIDGE

Cambridge University Press
978-0-521-00618-7 - Weighing the Odds: A Course in Probability and Statistics
David Williams
Frontmatter
More information

(e) It is possible that in future *quantum computers* will achieve things far beyond the scope of computers of traditional design. At the time of writing, however, only very primitive quantum computers have actually been made. It seems that (in addition to doing other important things) quantum computers could speed up some of the complex simulations done in some areas of Statistics, perhaps most notably those of 'Ising-model' type done in image reconstruction. In Quantum Computing, the elegant Mathematics which underlies both ANOVA and Quantum Theory may – and probably will – find very spectacular application. See Chapter 10 for a brief introduction to Quantum Computing and other (more interesting) things in Quantum Probability.

(f) My account of Statistics is more-or-less totally free from statements of technical conditions under which theorems hold; this is because it would generally be extremely clumsy even to state those conditions. I take the usual view that the results will hold in most practical situations.

In Probability, however, we generally know both exactly what conditions are necessary for a result to hold and exactly what goes wrong when those conditions fail. In acknowledgement of this, I have stated results precisely. The very concrete Two-Envelopes Problem, discussed at several stages of the book, spectacularly illustrates the need for clear understanding, as does the matter of when one can apply the so-useful Stopping-Time Principle.

On a first reading, you can (perhaps **should***) play down 'conditions'.* And you should be told now that *every set and every function you will ever see outside of research in Mathematics will be what is called 'Borel'* (definition at 45L) – except of course for the non-Borel set at Appendix A6, p499! So, you can – on a first reading – always ignore 'Borel' if you wish; I usually say 'for a nice (Borel) function', which you can interpret as 'for *any* function'. (But the 'Borel' qualification is necessary for full rigour. See the Banach–Tarski Paradox at 43B for the most mind-blowing example of what can go wrong.)

(g) Packages. The only package used extensively in the book is `WinBUGS`. See also the remark on `MLwiN` above. For Frequentist work, a few uses of `Minitab` are made. `Minitab` is widely used for teaching purposes. The favourite Frequentist package amongst academic statisticians is `S-PLUS`. The wonderful Free Software Foundation has made available a package `R` with several of the features of `S-PLUS` at [190].

(h) Note on the book's organization. Things in this-size text are more important than things in small text. All subsections indicated by highlighting are to be read, and all exercises similarly indicated are to be done. You should, of course, do absolutely every exercise. A ► adds extra emphasis to something important, a ►► to something very important, while a ►►► ... !

Do remember that 'the next section' refers to the next *section* not the next *sub*section. (A section is on average ten pages long.)

I wanted to avoid the 'decimal' numbering of theorems in the 'Theorem 10.2.14' style. So, in this book, 'equation 77(F1)' refers to equation F1 on page 77. If we were on page 77, that equation would be referred to merely as F1; but because of the way that LaTeX deals with pages, an equation without a page number could be on any page within one page of the current one. It's easier for you to cope with that than for me to tinker further with the LaTeX! Sometimes, the '□' symbol signifies a natural break-point other than the end of a proof. Forgive my writing 'etc' rather than 'etc.' throughout.

(i) Thanks. Much of this book was written at Bath, which is why that glorious city features in some exercises. My thanks to the Statistics (and Probability) group there for many helpful comments, and in particular to Chris Chatfield, Simon Harris, David Hobson, Chris Rogers and Andy Wood (now at Nottingham). Special thanks to Bill Browne (now at University College, London) and David Draper, both of whom, while MCMC enthusiasts, know a lot more than I do about Statistics generally. A large part of the book was written at Swansea, where the countryside around is even more glorious than that around Bath, and where Shaun Andrews, Roger Hindley and (especially) Aubrey Truman were of real help with the book.

Two initially-anonymous referees chosen by Cambridge University Press took their job very conscientiously indeed and submitted many very helpful comments which have improved the book. It is a pleasure to thank those I now know to be Nick Bingham and Michael Stein.

Anyone with any knowledge of Statistics will know the great debt I owe to Sir David Cox for his careful reading of the manuscript of the Statistics chapters and for his commenting on them with unsurpassed authority.

Richard Dawkins and Richard Tilney–Bassett prevented my misleading you on some topics in Genetics. Colin Evans, Chris Isham and Basil Hiley made helpful comments on Chapter 10.

In the light of all the expert advice, this really is an occasion where I really must stress that any remaining errors are, of course, mine. In particular, I must state that the book continued to evolve until after I felt that no further demands could be made on referees.

I typed the book in LaTeX: my thanks to Leslie Lamport and, especially, Donald Knuth. I am most grateful for willing help from Francis Burstall, Ryan Cheal and Andrew Swann (at Bath) and from Francis Clarke at Swansea in the cases where my LaTeX skill couldn't achieve the layout I wanted. Most of the diagrams I did in raw Adobe Postscript, some others with a 'C'-to-Postscript converter I wrote.

David Tranah of C.U.P. helped ensure that the book was actually written, suggested numerous improvements, and generally earned himself my strong recommendation to prospective authors. My thanks too to visionary artists, copy editors and other C.U.P. staff.

Malcolm and Pat, and Alun and Mair, helped me keep some semblance of sanity, and persuaded me that it was about time (for the sake of long-suffering Sheila – so worthy a chief dedicatee) that this book was finished. A big Thankyou to them.

For a fine rescue of my computer when all seemed to be lost, I thank Mike Beer, Robin O'Leary and, especially, Alun Evans.

For extremely skilled repair in 1992 on a machine in an even more desperate state, *me*, my most special thankyou of all to the team of miracle workers led by Dr Baglin at Addenbrooke's Hospital, Cambridge. It really is true that, but for them, this book would never have been written. Have a great new millennium, Addenbrooke's!

I am sad that the late great Henry Daniels will not see this attempt to make more widely known my lifelong interest in Statistics. We often discussed the subject when not engaged in the more important business of playing music by Beethoven and Brahms.

David Williams

Swansea, 2001

Please note that I use analysts', rather than algebraists', conventions, so

$$\mathbb{Z}^+ := \{0, 1, 2, \ldots\}, \qquad \mathbb{N} := \{1, 2, 3, \ldots\}.$$

($\mathbb{R}^+$ usually denotes $[0, \infty)$, after all, and $\mathbb{R}^{++}$ denotes $(0, \infty)$ in sensible notations.)