statistics books

I'd originally hoped to do a comprehensive survey of statistics books for this tutorial. For various reasons (that I'll describe below), this didn't quite happen as planned.

I first asked my friendly Prentice Hall rep. to send me any books intended for non-statisticians, titles like "statistics for social sciences" etc. For the next few weeks, almost every day, a parcel arrived! When the parcels stopped coming and my bookshelves were bending under the weight, I decided that one publisher was enough!

A complete list of these and other books I've looked at is at the end of this section, but the majority are from Prentice Hall! For general statistics books most publishers will have a similar range, so treat this as a 'random' sample.

The most interesting books seem to be the one-offs, special books not written in the general hackneyed style of statistics books. If you find any that you find particularly good for your own use or for students' use let me know and I'll add them to the tutorial web site.

classics

The classic book about the (accidental and deliberate) abuse of statistics is:

- **How to Lie with Statistics**, Darrell Huff. Penguin Books, London. 1991. (first published as **How to Lie with Statistics,** Darrell Huff and Irving Geis, Victor Gollancz, 1954)

This deals with both numerical and graphical ways in which statistics confuse, mislead and generally tell lies! It's also worth getting for Mel Calman's cartoons.

Of course, if you are producing graphs and want to do it well, then there is no better place to go than Tufte:

- **The Visual Display of Quantitative Information**, E.R. Tufte. Graphics Press, New York. 1983. (don't miss the 4,340 lb chicken on page 73)
- **Envisioning Information**, E.R. Tufte. Graphics Press, New York. 1990.

stats with no maths?

I was recently talking with a colleague who teaches statistics in a business school, who told me of a book which "teaches statistics without using any mathematics".

- **Statistics without Tears**, Derek Rowntree. Penguin Books.

Unfortunately, I haven't got myself a copy yet at the time I'm writing these notes, but will do so soon and update the web bibliography accordingly.

who's afraid of ...

For many an anxiety about mathematics starts in early, in the first years of school. There are various reasons:

- the progressive nature of the subject means that if once you loose track you've had it for ever
- the unforgiving precision of arithmetic leaves no room for getting better gradually
- the ability to follow rules and get the right answer without clear *understanding* means teachers may not realise there is a problem until too late
- and last, but by no means least, most teachers are themselves terrified of mathematics

So, perhaps it is at school rather than at an international academic conference that real effort is needed!

Even if we were to be able to put this right overnight, those who have left their school years behind can't simply forget this long established anxiety.

Among the books I've been looking at recently is:

- **Statistics for the Terrified**, Gerald Kranzler and Janet Moursund.
  Prentice Hall, Englewood Cliffs, NJ. 1995.

This starts from the assumption that the reader has just such a deep maths anxiety. The writer teaches statistics on counselling and education programmes and draws on both, even including an appendix on "Overcoming Math Anxiety"! The book is written in short punchy chapters, but does not patronise the reader and does include the statistical formulae (even if you're afraid of maths, you still need it!). I noticed few minor misunderstandings in the statistical details (page 15 says the wrong variance formulae minimises round off error and page 25 conflates two very different kinds of cumulative graph), but overall it looks good. In particular, chapter 7 gives a very good introduction to the principles of hypothesis testing and on page 10 there is an example of the difference between mean and median just like the one I give!

for the bold ...

In fact, those strong at mathematics often also find statistics difficult. They do all the formulae right, but miss out on the fact that the numbers all mean something! Statistics requires a blend of mathematics and the real world which is challenging and exciting. Quite like the conflicts and joys of working with computers and people!

Although the fundamental understanding of statistical concepts such as significance testing don't require deep mathematics (but do require deep thought!), the initial calculation of the formulae for these often do. This is especially important when new forms of statistical problem are encountered. For example, if you toss 20 coins you can look up in a standard table of the *binomial distribution* the probability, say, of getting 12 heads and 8 tails. However, recall the two horse races. You stopped when either heads or tails got to 10. What about the shorter column. How long do you expect that to be? Don't worry if you can't answer that quickly, it took me half an hour to work out and I won't bother to quote the formulae here. Suffice to say it isn't in the standard tables.

To help answer questions like this there is an extensive mathematical theory of statistics, some about particular kinds of statistical problem (for example time series analysis) and some about very general statistical properties (for example the central limit theorem that says nearly everything is approximately normal if its made of enough small bits). You don't need to understand any of this (nor 70% of what is in most general statistics books!) for day-to-day statistics. However, if you need to do something out of the ordinary, you need to get someone who can apply this general mathematical battery!

The two books that I refer back to for basic principles on this are:

- **Theoretical Statistics**, D. R. Cox and D. V. Hinkley. Chapman and Hall, London. 1974.
- **Mathematical Statistics 4e**, John E. Freund and Ronald E. Walpole Prentice Hall, Englewood Cliffs, NJ. 1987.

Did you notice that little word I slipped in ... 'basic' – do beware of any mathematics books entitled 'basic', 'introductory' or 'elementary' anything – they are usually advanced textbooks. In contrast, many that say 'advanced' or 'higher' are high school level!

general statistics textbooks

The general format of statistics books is rather similar ... a bit about graphs, a bit about probability and bit about significance testing. But there are differences in:

- level of detail – general statistics textbooks I looked at varied from 230 to 1182 pages!
- pedagogy – some in almost programmed learning style, others adopting a more 'how to do it' approach
- required prior knowledge – engineers are expected to know quite sophisticated mathematics before they start!
- examples – all the 'statistics for ...' textbooks tuned their examples to the relevant target audience.
- choice of subject matter – although broadly similar in content different domains employ specialised statistical tests or graphical displays (some I'd never seen before!), for example, both engineering and business books use 'control charts' to check that processes are running within specified limits, also some books spend more time on experimental method others on statistical modelling
- use of computers – many have some exercises targeted at computers uses, most commonly using MINITAB, but others rely entirely on hand computation; perhaps most surprising is only one book mentions use of spreadsheets (Excel) which for many people would be the obvious tool for small data sets.

cover story

Several hardback books I looked at put useful tables of formulae or statistical values on the inside front and back covers. Very useful for a book you buy yourself, but of course totally useless if the book is purchased by a library and has various labels stuck on its covers. Oh when will publishers realise that libraries need somewhere to put their labels!

finding things

I wondered how well various statistics books functioned as how-to-do-it guides, either as a reference or as a revision aid for students. As a test case I tried to look up regression in each assuming I wanted to work out the line of best fit through some data points. To make things easy I assumed I knew enough that it was the 'regression' equation I was after – I didn't try looking up line (just as well, the few I looked at the word didn't appear).

First thing. In none of the books I looked at was there any indication in the index of the defining location within the book. The closest I found was in.

- **Elementary Statistics 2e**, Neil Weiss. Addison Wesley, Reading, MA. 1993.

Although the ordinary index doesn't show defining references, there is a handy 'procedure index' on the inside back cover (yes the cover!). This has part labelled 'regression' – I quote:

> Slope of the population regression line
> Confidence intervals, 639
> Hypothesis tests, 636

Oops, no entry for the actual calculation of the slope!

In this and nearly all cases the best course of action was to find the relevant chapter through the table of contents and to skim the chapter until I found the formulae.

In fact, this book also has a pull-out reference card which includes the regression formula, but again, alas, no reference back to the text.

(Incidentally, I have had this book for several years since an Addison Wesley rep. called round and assured me she had a really good book on statistics. Unfortunately, I found its rather noisy two colour scheme and programmed learning approach almost impossible to read. However, this is very much a matter of taste and looking at it again during this review process I found its mini-biographies and motivating case studies very engaging.)

Full marks to:

- **The Essence of Statistics for Business 2e**, Michael C. Fleming and Joseph G. Nellis. Prentice Hall, Hemel Hempstead 1996.

In the index I found:

> regression equation 183, 197

Joy of joys, page 183 has the actual equation:

> Formulae for regression coefficients:
>
> $$b = \frac{\Sigma X_i Y_i - n\,\overline{X}\;\overline{Y}}{\Sigma X_i^2 - n\,\overline{X}^{\,2}}$$
>
> $$a = \overline{Y} - b\,\overline{X}$$
>
> where $\overline{Y}$ and $\overline{X}$ are the arithmetic means of the Y_i and X_i values respectively

Furthermore, page 197 is the chapter-end summary of 'Key learning points' which lists all the major terms and equations in the chapter.

Second prize goes to:

- **Statistics for the Social Sciences** Victoria L. Mantzopoulos. Prentice Hall, Englewood Cliffs, NJ. 1995.

This book has a lovely mini-dictionary/glossary in appendix B. Unfortunately the glossary doesn't refer back to key pages ion the book! Never mind, I turn to the index:

Regression, 350–51, 370
  bi-variate, 64–67, 69
  multiple, 67–69

I wisely ignore the two references late in the book (actually to a case study and the mini-dictionary entry!) and plumb for 'bi-variate'. Bingo! The formulae for the regression line and how to calculate the various coefficients in three grey boxes.

Sadly this is the best there is. A more typical example was:

- **Basic Statistics for the Social and Behavioural Sciences** George M. Diekhoff. Prentice Hall, Englewood Cliffs, NJ. 1996.

A nicely presented book with good use of margins and two colour to highlight key definitions and equations. It has a glossary, but with no cross references to the text! It also has 'Essential Equations' listed on inside front and back covers (!!). On the inside back cover I find:

MAKING PREDICTIONS

Eq.11.1 $Y' = a + b_Y X$ where $b_Y = r\left(\frac{s_Y}{s_X}\right)$

$a = \overline{Y} - b_Y \overline{X}$

On the inside front cover I find the definition of '$\overline{X}$' and of 's', and, let's say, I remember that 's_Y' and 's_Y' are applications of the same formula as 's' with values of Y and X respectively. But what is 'r'? A few lines above there is a definition of 'r_p', which has nothing to do with it! (I know this, does the typical reader?)

I turn to the index and after a few false starts find equation 11.2 which is the formula for b_Y. (Incidentally 11.1 is just $Y' = a + b_Y X$.) This also has the ellusive 'r' but also an explanation:

b_Y = the correlation between X and Y

At last I'm on to something. I look up 'correlation' in the index. This has some main entries and lots of sub-entries. The main entries are no good and rather than scanning all the subentries, I look back at the text around equation 11.2. Just below, in the midst of the text I read 'multiply the resulting value by the Pearson correlation between X and Y (r)' – at last!! I look up 'Pearson correlation' and get there.

Note that this doesn't mean that Diekhoff is particularly bad, in fact overall it looks pretty good (even mentions squared dollars!), it's just that books in general tend to have pretty poor indices and statistics books are no exception. Unfortunately this makes them pretty useless when you want to use them as a how-to-do-it reference.

tables

Most textbooks include some statistical tables in an appendix. For more detail you may want to get a special book of statistical tables. The one I use is:

- **Statistical Tables**, F.D.J. Dunstan, A.B.J. Nix and J.F. Reynolds. R.N.D Publications, Cardiff. ISBN 0-9506719-0-8. 1979.

Another popular book of tables is:

- **Elementary Statistical Tables**, Henry R. Neave. Allen & Unwin, London.

I don't know if this is a problem in other countries, but British students now have great difficulty reading tables because the use of calculators has meant that old logarithmic and trigonometric tables are no longer used in school. For students' use I would recommend avoiding any books of tables that have 'interpolation columns' as these are particularly confusing (albeit useful).

Also, a word of warning, be very careful with:

- **value → significance vs. significance → value.** Some tables are listed by significance value, often at 5%, 2.5%, 1% and 0.5%. You look up the significance you are after, say 5%, and then the table tells you the 'critical value', that is if your calculated value is bigger than this critical value it is 5% significant. Other tables, notably tables of the Normal distribution and also discrete distributions such as the binomial work the other way round. You look up your value in the table and then it tells you the probability (significance) of that value.
- **probability or percentage.** Tables may use probabilities (e.g. 0.05) or percentages (e.g. 5%) and may also list values by the probability that a value is less than a value, which may mean you have to look up the 0.95 (probability) or 95% (percentage) point.
- **one- and two-tailed tables.** First you need to know which you want – no easy matter even when you know what the words mean. Then you have to look it up... For asymmetric distributions (e.g. $\aleph^2$) both tables may be given (or combined in a single dual purpose table), for symmetric distributions one or other may be given in the tables and you are expected to work out the other (e.g. for Students' t distribution to find the two-tailed 5% point you look up the 2.5% one-tailed value from the tables!). Worst of all is the F distribution. Most frequently this is used in ANOVA tests and the tables are hence given in one-tailed form. However, occasionally you need the two-tailed version and then you have to work out a weird formulae including the reciprocal of the value in the table (Just hope you never need that one!).
- **Normal distribution.** These are usually given in one-tailed form (although sometimes two-tailed as well). But they have at least three different forms. (i) First you have tables which give the one-tailed significance value – that is the probability that a value is greater than the given value). (ii) Others give the probability that the value is less than the quoted value. (iii) Finally some give the probability of a value between zero and the given value. If you looked up the value 1.65 in these tables you will find: (i) 0.05 (ii) 0.95

(iii) 0.45. Note that the Normal table in 'Statistics for the Terrified' gives both (iii) and (i) side by side.

Many statistics packages still expect you to look up the value in a table, so this, possibly one of the most confusing aspects of statistical computation, is not always eased by the use of computers.

book list

Here is the complete list of books that I've considered. A preponderance of Prentice Hall titles for the reasons I gave earlier! Do tell me about other books you find useful, especially those that address conceptual understanding.

- **How to Lie with Statistics**, Darrell Huff. Penguin Books, London. 1991. (first published as **How to Lie with Statistics,** Darrell Huff and Irving Geis, Victor Gollancz, 1954) (124 pages)
  • I've been meaning to get my own copy of this for years and eventually did ready for this tutorial. Well worth it!
- **The Visual Display of Quantitative Information**, E.R. Tufte. Graphics Press, New York. 1983.
  • The classic book of graphical representation of data. Don't miss the 4,340 lb chicken on page 73.
- **Envisioning Information**, E.R. Tufte. Graphics Press, New York. 1990.
  • Second book by Tufte. Also well worth reading. Both are a visual delight!
- **Statistics without Tears**, Derek Rowntree. Penguin Books.
  • Going on third party recommendation here, will try to get hold of a copy soon.
- **Statistics for the Terrified**, Gerald Kranzler and Janet Moursund. Prentice Hall, Englewood Cliffs, NJ. 1995. (164 pages)
  • Interesting to note how this book addresses its aims: (i) short chapters (ii) lots of examples (iii) personal language (iv) doesn't attempt to give reasons for formulae.
  • I liked chapter 7 'introduction to inferential statistics' which gives concepts first.
  • There are one or two minor problems (see 'who's afraid of ...' section above). • Also interesting to note that the appendix on 'overcoming math anxiety' is mainly about general techniques to deal with debilitating anxiety, but doesn't take the chance to make the solutions specific for mathematics. It seems we all have trouble when too close to our own areas of expertise.

Tables:

- **Statistical Tables**, F.D.J. Dunstan, A.B.J. Nix and J.F. Reynolds. R.N.D Publications, Cardiff. ISBN 0-9506719-0-8. 1979. (67 pages)
- **Elementary Statistical Tables**, Henry R. Neave. Allen & Unwin, London.

General (as opposed to 'statistics for ...') statistics textbook:

- **Elementary Statistics 2e**, Neil Weiss. Addison Wesley, Reading, MA. 1993. (733 pages)
  • Good double page chapter start spreads including a mini-biography (I love them!) and the introduction of a case study. The case study is revisited at the end of the chapter and 'solved' using the techniques introduced givings a sense of purpose. • Chapter ends also good with review of key terms and formulae and exercises, including some using MINITAB. • The book has a bit of a programmed learning feel which I've never got on with personally. • Great reference card (even better if it referenced text pages). • Useful, (but incomplete) procedure index on inside back cover. • Example data sets listed in an appendix and used in exercises.

Now the 'statistics for ...' books from Prentice Hall

- **The Essence of Statistics for Business 2e**, Michael C. Fleming and Joseph G. Nellis. Prentice Hall, Hemel Hempstead 1996. (270 pages)
  • Won hands down in my 'look up regression' test. • 'Use of MINITAB' section at end of each chapter. • Short punchy (perhaps not so easy?) treatment with plenty of diagrams and examples. • 'Key learning point' section at end of each chapter summarising main definitions and equations.
- **Statistics for the Social Sciences** Victoria L. Mantzopoulos. Prentice Hall, Englewood Cliffs, NJ. 1995. (382 pages)
  • Second in my 'look up regression' test. • Separate study guide (I haven't seen it) includes use of SPSS and MYSTAT. • Grey boxes for key formulae and definitions. • List of hypothesis testing formulae in appendix (not regression!). • Dictionary of terms and formulae in appendix (but no reference back into the text!). • Ignore the shape of the Chi squared curves in chapter 12 – the graphic artist obviously didn't know any statistics!
- **Basic Statistics for the Social and Behavioural Sciences** George M. Diekhoff. Prentice Hall, Englewood Cliffs, NJ. 1996. (448 pages)
  • Despite my tale of woe looking up regression formula, it is quite a nicely laid out book with good use of two colour and lots of good features. • Great introduction giving an overview of all the key statistical ideas in the book. • Pink boxes for key formulae and definitions. • Keypoints highlighted in sidebars. • Front and back covers used to list key formulae (no page number reference back into the text, but equation numbers given.) • Separate SPSS workbook. • Appendix dedicated to explaining use of summation symbol. • Worked examples at end of book. • Glossary – but no page references!
- **A First Course in Business Statistics 6e,** James T. McClave and P. George Benson. Prentice Hall, Englewood Cliffs, NJ. 1995. (746 pages)
  • Full data set available separately on floppy. • Generic (i.e. program independent) 'Using the Computer' exercises at the end of each chapter (also section 10.8 'Using the Computer for Regression' telling you how to do it in SAS). • Many SAS and some SPSS printouts used in text. • Floppy disk enclosed with book with ASP the student version of a statistical package ... sounds good? Unfortunately, although there is a getting started ASP tutorial in an appendix there is no other mention in the book. • (Very) brief mention of Baysian statistics.

Books for scientists and engineers. Note that engineers and 'hard' scientists are assumed to have more initial mathematics.

- **Statistics for the Biosciences**, William P. Gardiner. Prentice Hall, Hemel Hempstead 1997. (416 pages)
  • Subtitled 'data analysis using minitab software', the book does use MINITAB extensively throughout. • Appendix of 'statistical formulae' ... you guessed it, no reference back to page numbers in the text!. • Every chapter ends '... in the biosciences', or '... for biological experimentation', so you can't forget what book you're reading. • Lots on experimental design.

- **Introductory Statistics for Environmentalists**, Paul Moore and John Cobby. Prentice Hall, Hemel Hempstead. 1998. (250 pages)
  • Useful chapter on 'survey methods', which is not (as one might imagine) about transepts of mud flats, but actually about ideas of different kinds of variable, random sampling etc. • Back cover says 'One of the few texts to make reference to MINITAB for Windows and Excel for Windows'. Although it makes more mention of Excel than any other book I looked at (not difficult), you'd better not blink.

- **Statistics for Engineering and the Sciences 4e,** William Mendenhall and Terry Sincich. Prentice Hall, Englewood Cliffs, NJ. 1995. (1182 pages)
  • Second book with ASP floppy (did Prentice Hall get a job lot cheap?) ... this time no mention whatsoever of ASP software in text with exception of 'supplements' list in preface which says the floppy is included. • Computer lab section at the end of each chapter giving work-through in SAS and MINITAB. • Large data sets listed in appendix and available separately on floppy (why not include this rather than ASP!). • Brief mention of Baysian statistics.

- **Miller and Freund's Probability and Statistics for Engineers 5e,** Richard A. Johnson. Prentice Hall, Englewood Cliffs, NJ. 1994. (630 pages)
  • Some MINITAB exercises. • Chapter on reliability and testing • Special tables for engineering data (e.g. sample size code letters from MIL-STD-105D) • Short example of Baysian statistics.

- **Statistics for Analytical Chemistry 3e**, J. C. Miller and J. N. Miller. Ellis Horwood, Chichester. 1994. (233 pages)
  • Assumes good basic knowledge of probability and maths. • Treatment of issues such as outliers in data and propagation of experimental errors during subsequent computation. • Lots of specialised issues: control charts, use of regression for calibration. • Last chapter includes use of optimisation techniques for finding optimal factors and pattern recognition techniques.

Books I've used myself – heavy mathematics!

- **Mathematical Statistics 4e**, John E. Freund and Ronald E. Walpole Prentice Hall, Englewood Cliffs, NJ. 1987. (511 pages)
  • Mathematical derivations of many common statistics.

- **Theoretical Statistics**, D. R. Cox and D. V. Hinkley. Chapman and Hall, London. 1974. (608 pages)
  • More fundamental again, doesn't deal with specific tests and statistics, but general principles such as least-squares estimates, Baysian methods etc..