

[End of previous article - see below]

Title: Foundations of Statistical Analyses and Applications with SAS
Author: Michael Flak, Frank Marohn, Bernward Tewes
Birkhauser Verlag, Basel (2002) ISBN 3-7643-6893-4

Reviewer: John Norrie
Email: j.norrie@stats.gla.ac.uk

This ambitious book attempts two considerable tasks – to give the foundations of statistical analysis, and apply these foundations using SAS. On one hand, SAS is a massive, complex environment, taking years to learn and use efficiently. On the other hand, to successfully introduce the foundations of statistical analyses is probably several books in itself.

The authors indicate the audience is primarily students of statistics and mathematics, and other disciplines (economics, biostatistics) where statistics has a strong impact. They also see a use for the practitioner who, beyond using statistical tools, is interested in their mathematical background. The book assumes a “higher level of mathematics than [for] most applied statistics books”, and is targetted at students who have “already had an introduction to probability theory and mathematical statistics”. It does not assume any SAS knowledge. Their motivation is to bridge the gap between the theory and practice – since analysis of real data using statistical methods with a software package common in industry is not usually an integral part of mathematical statistical studies.

In terms of teaching, the book is designed as a two semester course, with the first four chapters (with a strong theory component) in semester one, and the remaining 5 chapters (more applied, in particular multivariate techniques) in the second semester, delivered as lectures, practical training, and tutorials. This book would not be suitable for an introductory or service course. It may be suitable for a more advanced undergraduate course, but its real place would be found

in a postgraduate course attempting a mathematically rigorous introduction to statistical analyses and exposing the student to mainstream statistical programming concepts and data management tools (in SAS). It might be useful as a self teaching tool, possibly as an adjunct to a course in mathematical statistics.

Throughout the book the SAS examples are delivered as a program code box plus explanatory notes and perhaps data listings, with an output box containing the results. The programs and datafiles are available at <http://statistics-with-sas.ku-eichstaett.de>. In general, the datasets are interesting, covering a wide range of studies (e.g. the prognostic importance of pre- and extra-marital sex in divorce). It is easy to download from the website, but no instruction is given on how to extract the data (for example, the requirement to change the directories to suit a local machine may cause confusion).

In terms of coverage, there are 8 chapters and an Appendix (A Brief Introduction to SAS). The first four chapters are the mathematical spine of the book - Elements of Exploratory Data Analysis, Some Mathematical Statistics for the Normal Distribution, Regression Analysis, and Categorical Data Analysis.

The next four chapters, the applied material, which the authors recommend selecting from, cover Analysis of Variance, Discriminant Analysis, Cluster Analysis, and Principal Components.

Each chapter contains many Exercises, with a good mixture of mathematical set pieces – proofs and demonstrations of key results – and applied exercises using SAS code on real data. However, disappointingly no solutions or SAS output is given. Such detail would have been useful – if say, available on the data-hosting website, or as a Teaching Pack to aid the class instructor.

Chapter 1 (Elements of Exploratory Data Analysis) introduces some graphical tools for visualising data, with an interesting treatment of histograms and kernel density estimators. Stem and leaf plots follow, with discussion of robust measures of location and spread, boxplots and quantile plots, and novel material on hanging histograms and rootograms as goodness-of-fit tools. There is discussion on variance stabilising transformations, and later (page 72) a section on log-transforming data to better meet a Normality assumption. A criticism would be the data transforming material is purely theoretical, with no mention of the consequences of, or appropriateness of, such transformation in particular contexts (e.g. is a multiplicative scale meaningful?). PROC FREQ is explored, plus basic graphics procedures (GPLOT and GCHART), along with the relatively new PROC BOXPLOT. Several key data manipulation procedures are covered (including PROC SORT). The SAS concept of the 'Data Step' for reading in and manipulating SAS datafiles is introduced, along with DO statements - giving the novice instruction in data array processing.

Chapter 2 covers Some Mathematical Statistics for the Normal Distribution, and contains the usual detail on Normal, Chi-Squared, F and t distributions. There are brief details of the concepts of hypothesis tests and confidence intervals, followed by the two sample t-test and the 1-sample paired t-test. The Wilcoxon rank sum test is presented with a useful section on the treatment of tied data. The summarising procedures in SAS illustrate the points made – PROC MEANS and UNIVARIATE, with also the testing procedures PROC TTEST and NPAR1WAY.

Chapter 3 introduces linear regression analysis, with a routine exposition of simple regression, the method of least squares, and on to multiple regression. There is a section on polynomial regression which, given the competition for space, seems a slightly odd choice. The relevant statistical procedures are illustrated (e.g. PROC REG) and enhanced graphics given by using 'annotate'

options in PROC GREPLAY for saving and reshaping graph catalogs.

Interestingly, SAS %macros are introduced by invoking a supplied macro %mkfields. I would have either left the macro facility out altogether or introduced this very powerful but difficult to use SAS capability in a more detailed fashion. Also, the reader sees for the first time the MERGE statement, a cornerstone of the SAS data processing – but without a BY statement for matched-merging, an inherently dangerous practice. Elsewhere it is claimed that SAS/FSP is a "collection of procedures such as FSEDIT that facilitates SAS data handling". Whilst true, this target audience might be better learning about data handling via DATA steps. Similarly PROC TRANSPOSE, a very important data manipulation tool, gets only a fleeting and somewhat convoluted mention (page 186).

Chapter 4 continues with an exposition on Categorical Data Analysis. There is a pleasing section on Fisher's Exact Test, and the authors make good use of SAS data step, graphical procedures, and SAS mathematical functions to illustrate the properties of the hypergeometric distribution. There is very brief mention of Fisher's test being a conditional test, and in a teaching context it is pity this is not developed further. I found the section on Categorical Regression hard going – and the illustration of a logit model using PROC CATMOD positively baffling. I was not sure why the authors ignored the much easier to use PROC LOGISTIC or the even newer PROC GENMOD. There is a brief mention of neural networks, and survival analysis via Cox regression is given a short sentence.

The remaining chapters – Analysis of Variance, Discriminant Analysis, Cluster Analysis, and Principal Components – give readable accounts of these various analysis techniques. In the Analysis of Variance chapter, the vexed issue of multiple comparisons is given an inadequate one-liner (page 198 "But recall that you have to fix the error level beforehand"). In terms of teaching, one either has to carefully explain the issue or instead leave it for another day - such a statement will excite but not satisfy the curiosity of a discerning student. By page 216, the authors seemed to have temporarily thrown in the towel with the exposition of the SAS procedures, with a reference to PROC TABULATE reading "cf SAS Procedures Guide (SAS OnlineDoc)". I'm not criticising them for this – why re-invent the wheel? – but it would be better if SAS OnLineDoc was properly introduced early on. There are good sections on the use of PROC ANOVA and the Kruskal-Wallis test, and as with all the rest of the book, the mathematical development is readable and thorough.

I found the chapter on Discriminant Analysis the most interesting, probably because it was the topic I knew least about. Interestingly, on page 256 they neatly introduce PROC IML, the SAS interactive matrix language, so the user sees immediately by example the possibilities of IML. PROC DISCRIM is well covered, and PROC G3D used to illustrate SAS 3-dimensional graphing capabilities.

The next Chapter on Cluster Analysis was likewise of interest, and by now the examples are using most of what has been previously introduced – SAS data processing procedures, SAS Data Steps, IML, SAS analysis procedures, and SAS Graphics tools (for example, in their exposition on Multidimensional Scaling), with PROC CLUSTER and PROC TREE getting a useful airing.

The last Chapter concludes with material on Principal Components. There is a pleasing geometrical exposition contrasting principal components with the geometry of linear regression. PROC PRINCOMP is covered, and then the authors develop the topic into factor analysis techniques using PROC FACTOR.

There is a brief introduction to SAS as an Appendix. To introduce SAS in 17 pages would be an heroic feat. Although not assuming any SAS knowledge, the authors state that “the level of SAS programming should be no serious problem for a student of maths or stats who has some practical knowledge of an arbitrary software package”. Well, yes and no. SAS, although it has the essential programming logic of any language, is unique, and can be very frustrating. It is in essence a batch-submit programme, not an interactive dynamic environment. Why then have the authors chosen SAS? The stated aim is to bridge the gap between theory and practice incorporating a widely used non-academic computer package. But why not S-Plus, or Minitab, or SPSS even? No mention is made of alternatives, which is a pity.

Before getting to the conclusion, there were a number of other issues. First, this is a translation - originally published in German under the title “Angewandte Statistik mit SAS, Eine Einführung”, Springer Verlag, 1995. As with all translations, there is scope for interpretation, and this book will doubtless read better in the original German. Overall, the translation is readable, but the sentences can be somewhat mysterious on occasions. Second, there are a number of typos (eg page 180 “Titterton” instead of “Titterington”) but more irritatingly in several places the SAS examples just end abruptly (clearly text is missing – e.g. page 187 Program 5_1_2). Third, the SAS material was on the whole up to date, with the 1995 edition clearly having been revised, so the

English edition gives examples in SAS 8.

So, in conclusion, although my initial scepticism that a single book could usefully contribute to both its stated aims of providing both the mathematical background to statistical analyses and the means to perform examples of such analyses using SAS was in part borne out, nevertheless this is a worthy attempt. If we then consider that any book that attempts such feats is bound to fail in some sense, what is the quality of the failure here? At just under 400 pages, it is commendably brief. On balance, I liked this book. Yes, some of the omissions were surprising, and some of the topics covered in one-liners raised an eyebrow (e.g. data transformations, multiple testing, survival analysis). All instructors would probably chose different material and different styles to get novice students from nowhere to basic competence in the SAS environment, but I think the authors probably achieve this for their target audience of mathematical statisticians. As a mathematical statistician who has used SAS for a dozen years, I found the SAS examples they gave generally informative, even throwing up some unexpected “nuggets”. I found the mathematical treatment well organised and usually insightful. There are many good SAS books (even more not so good) - two of direct interest are those by Der and Everitt – a very usable and comprehensive introduction, and the teach yourself style of Aster’s book on SAS programming – which can serve both a teaching tool and a reference. Excellent introductory texts on the foundations of statistical analysis include (from the biostatistical literature) Steel and Torrie, Armitage and Berry, and Lloyd Fisher. In a very competitive market with established teaching materials for both aims, whilst I would recommend this book as a useful addition to the teaching arsenal, it is not a one-stop solution. However, as the authors wisely say in their introduction “*Statistics is never having to say you’re certain*”, and that should equally apply to this review...

References

- [1] Steel R and Torrie J, *Principles and Procedures of Statistics – A Biometrical Approach, 2nd Edition*. McGraw-Hill, 1980
- [2] Fisher L D and Van Belle G, *Biostatistics – A Methodology for the Health Sciences*. Wiley, 1993
- [3] Armitage P, Berry G and Matthews J N S, *Statistical Methods in Medical Research, 4th Edition*. Blackwell, 2002
- [4] Der G and Everitt B S, *A handbook of Statistical Analyses using SAS, 2nd Ed*. CRC Press, 2001
- [5] Aster R, *Professional SAS Programming Logic*. Breakfast Communications Corporation, 2000