Journal of Official Statistics, Vol. 14, No. 1, 1998, pp. 97–111

Book and Software Reviews

Books for review are to be sent to the Book Review Editor Jaki Stanley, USDA/NASS, Research Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.

Leon Willenborg and Ton de Waal. Statistical Disclosure Control in Practice. Heidelberg: Springer-Verlag, 1996. ISBN 0-387-94722-1. 142pp+ refs and index. 54 DM.

The importance of statistical disclosure control (SDC) in official statistics, and also for data reporting in many other areas, is undoubted. However, a significant proportion of the available information about the motivations, techniques, and methodologies for SDC could be seen as somewhat difficult to access for some practicing statisticians and researchers. Much excellent work has unfortunately appeared only in reports and conferences.

Given the lack of publications providing complete coverage of the field from an introductory level and the quantity and quality of research not easily accessible to all those who may be in need, there would appear to be a definite need for a number of books on the subject of SDC. Ideally, these would provide reasonably complete and contemporary overviews of various aspects of the field. This volume attempts to provide such an overview for practicing official statisticians who need to deal with disclosure control issues and is therefore a very welcome addition to the field.

In the preface the authors point out that the book is not written from a statistical/mathematical or legal perspective, but there is still a reasonable amount of material contained within that would be of interest to both these groups, and others who are not official statisticians. However, the focus is very much from the perspective of an official statistician and the selection of topics is obviously influenced by the authors' association with Statistics Netherlands.

The book contains eight chapters, the first of which provides an introduction to the motivations and concepts behind statistical disclosure control. This chapter introduces the importance of the field, and provides a brief introduction to global recoding, local

suppression, and substitution techniques for microdata. Disclosure control for tables is then covered, with similarities to microdata noted.

The second chapter expands on the first with more detailed discussion of the concept of disclosure. Disclosure is here covered in a more formal manner, building on the intuitive explanations already presented. This chapter continues the trend started in the first of discussing microdata and tabular data in turn.

Chapter three provides an overview of SDC policy for several national statistics organizations. Various options for data dissemination are listed and rated in terms of their information content, the need for licensing, and the use of on-site work. Following this, several countries' official statistics organization are briefly discussed. The countries covered are Australia, Canada, Denmark, France, the German Federal Republic, Great Britain, Italy, the Netherlands, New Zealand, Norway, Sweden, and the United States. The Netherlands and Great Britain are then examined in more detail as case studies. The information presented here was interesting, albeit much too brief, and this chapter could have been considerably lengthened with more detail on some of the countries' use of SDC.

Chapters four and five specifically look at SDC for microdata, with Chapter four making up the discussion section. Issues covered in this chapter include the different types of variables (identifying/non-identifying, sensitive, household, and regional) and the rarity of values and combinations. Some examples of SDC rules are presented which could be useful for official statisticians needing to formulate their own set of rules. Chapter five then examines the mathematics of some of the specifics of microdata disclosure control in more detail.

The same pattern is repeated in Chapters six and seven for tabular data. Again, Chapter six provides a discussion of issues that arise and some SDC rules. This chapter makes extensive use of examples to illustrate the disclosure risk of tables and the disclosure control strategies, including some discussion of linked tables. Some practical recommendations are made which could again be useful for the statisticians formulating their own procedures. This is followed by a mathematical treatment of some areas in Chapter seven.

The book concludes in Chapter eight with some discussion of software issues and then short sections on some of the areas mentioned in the text that the authors feel justify more research and some newer areas that the book does not otherwise mention. Substitution techniques are briefly discussed here, and this is one topic that should have been covered in much more detail within the main body of the text.

The small size of the book is a double-edged sword; a larger volume may have been impractical and off-putting to some practitioners, but the volume's short size means that many topics are, at least partially, neglected. More attention to substitution techniques would have been especially welcome. While the authors' tendencies are very much towards suppression and recoding techniques, much useful work has been carried out on substitution methods and this deserved considerably more mention than the passing references. Another aspect that could be seen as lacking is the effect of SDC from an analyst's perspective. The audience of the book could have been considerably widened, and the perspective of the official statistician enhanced, by more treatment of the effects of SDC on statisticians using the data. Some reference is made within the text to information loss, and this is also a topic mentioned in the concluding chapter, but perhaps this topic

deserved its own chapter as well. However, this omission as with substitution techniques mentioned above, merely reflects the flavor and approach taken for the book.

The references contain many reports (22 of the 79 references are reports from Statistics Netherlands) and conference papers that may be difficult for some practitioners and researchers to obtain easily. Given the small size of the volume, more accessible references would have allowed interested readers to further explore areas of interest. An up-to-date annotated bibliography would also have helped to make readers more aware of the literature and options available.

The index could also be much improved (for example there is no entry for 'noise', but this is found under 'adding noise'), with more entries warranted despite the book's small size. Some of the writing style could also have been refined to make the contents flow better. Occasionally the book repeats the same information in a very similar manner in more than one section, and while the repetition may not be harmful it can be distracting. A small number of errors may catch the unwary reader but should not cause too many problems for the book's intended audience.

Despite these criticisms the book provides a valuable starting point for anyone who currently deals with, or may in the future have to deal with, statistical disclosure control issues, or is interested in why the data they obtain has been altered to protect against disclosure. The current edition could readily be expanded into a standard text with the addition of some of the areas mentioned above.

Andrew Gray
Software Metrics Research Laboratory
Department of Information Science
University of Otago
New Zealand
Phone: +64 3 479 5282
Fax: +64 3 479 8311
email: agray@commerce.otago.ac.nz
http://divcom.otago.ac.nz:800/com/infosci/smrl/home.htm

Graham R. Walden. Polling and Survey Research Methods 1935–1979: An Annotated Bibliography. Greenwood Press, Westport, Connecticut and London, 1996. ISBN 0-313-27790-7. Xxx+581pp. 99.50 USD.

In 1990, Graham R. Walden published a selective, annotated bibliography of survey methodology and public opinion polling, compiling summaries of 359 scholarly books and articles published in the U.S.A. in the 1980s (Walden 1990). This book extends Walden's prior efforts by summarizing 1,013 scholarly books and articles published between 1935 and 1979, again limited to the U.S.A.

The book's preface explains its scope, coverage and source materials. It is organized around 17 sections, including reference sources, instructional materials, history, pollsters and polling organizations, overview studies, design and planning, sampling, questions, interviewers, interviewing, mixed mode data collection methods, respondents, responses, analysis, discipline-oriented studies and applications to specific areas, special topics, and

humor. Each section is further divided in up to 33 labeled sub-sections, ranging from interviewing Navajos, weighting data, ethics, response latency, use of tape-recorders, and polls for each presidential election from 1936 to 1968.

Each of the 1,013 books and articles reviewed is numbered, to allow easy cross-referencing. Each summary includes a concise statement of the problem or purpose, data used and conclusions, as well as counts of footnotes and references. The average summary is one-quarter to one-half page. Since any given book or article may naturally cover more than one topic, Walden assiduously refers readers to appropriate cross-references by the review number. Not surprisingly, the plurality of summaries come from the primary journal for methodological and theoretical research in the field, Public Opinion Quarterly. A substantial number of articles also come from the Journal of the American Statistical Association, the Journal of Marketing Research, the Journal of Applied Psychology, and the Journal of Advertising Research. The inclusion of articles from tens of more obscure journals is further testament to the comprehensiveness of this bibliographic reference.

The body of the book is followed by four appendices, covering acronyms, source journals, print and CD-ROM sources, and survey research professional organizations. The book concludes with three indexes, covering authors, selective keyword stop words, and selective keywords.

The sheer magnitude, tight organization, and encyclopedic appendices and indexes of this book demonstrate that its compilation ''has clearly been a labor of love,'' as Seymour Sudman remarks in his introduction. Walden intends for the volume to provide ''the reader with bibliographic access to the first forty-five years of polling and survey research utilizing scientific sampling.'' And indeed it does. Readers can easily scan the 450 pages of reviews for those most appropriate to their interest. Moreover, the organization of this historical volume closely parallels Walden's previous book of survey research and polling summaries in the 1980s, easily allowing cross-volume examination.

Beyond bibliographic reference, however, Sudman's introduction points out that this volume will be essential to historians and social scientists interested in the evolution and expansion of scientific survey research and polling over the past sixty years. Sociologists of science, for example, could use this volume in conjunction with Walden's previous book to uncover the timing and sequence of methodological innovations and adoption, in survey sampling, question construction, data collection procedures, interviewer training, and data analysis in the U.S.A. (see, however, Converse 1987, for prior research in this area). Together, the companion volumes also allow sociologists of science to uncover the principle players' productivity and influence in the development of scholarly and applied survey research and public opinion polling. The only criticism which might be leveled, mentioned by Sudman, is that the book omits the contributions of survey innovators in other countries.

While this volume is not likely to be found on the bookshelves of the average survey research practitioner or survey data collection facility, it is essential for any comprehensive academic research library.

References

Converse, J. (1987). Survey Research in the United States: Roots and Emergence 1890–1960. Berkeley, CA: University of California Press.

Walden, G.R. (1990). Public Opinion Polls and Survey Research: A Selective Annotated Bibliography of U.S. Guides and Studies from the 1980s. New York and London: Garland Publishing, Inc.

Patricia A. Gwartney
Oregon Survey Research Laboratory
and Department of Sociology
University of Oregon
Eugene, OR 97403-1291, U.S.A.
E-mail: pattygg@oregon.uoregon.edu
http://darkwing.uoregon.edu/~osrl
Telephone: (541) 346-5007
Facsimilie: (541) 346-5026

C.G.G. Aitken. Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley and Sons, Inc., Chichester, West Sussex, England, 1995. ISBN 0-471-95532-9. 260pp (cloth). £34.95.

Locard's Principle states that every contact leaves a trace. Aitken expands on this principle by providing statistical models with which we can estimate the probability of the presence of certain trace evidence on a suspect or at a crime scene if a suspect is guilty. The goal of the book is not to determine whether individual suspects are guilty or innocent, but rather to allow scientists to assess the value of given pieces of evidence. Judges and/or jurors, then, can use the estimates to aid them in their decisions about guilt.

Aitken begins his book with a brief discussion of uncertainty and the role it plays in science. He presents the reader with an introduction to probability, beginning with traditional die rolling and card drawing examples, then progressing into examples concerning blood types and paternity suits. He also illustrates how probabilities can be updated as more evidence becomes available.

The second chapter is devoted to a review of the Bayesian approach in general and presentation of the odds form of Bayes' Theorem, which allows scientists to determine the ratio of the probability of guilt to the probability of innocence, given the presence of certain evidence. The chapter also discusses two common errors in forensic statistics: *prosecutor's fallacy* and *defender's fallacy*. Both fallacies result from viewing the probability equation through a biased lens. Hence, prosecutors may argue that if a rare trait found at a crime scene is found in only 1% of the population, there is a 99% chance that a suspect who possesses that trait is guilty. Of course, this is untrue; one also needs to know the probability of the trait occurring if the suspect is innocent. Defenders may counter with their own fallacy by stating that 1% of the population in a large city may mean that several thousand citizens possess the rare trait; thus, the evidence ''has little relevance'' because the suspect is one of thousands who may have committed the crime. While this argument is not technically untrue, *defender's fallacy* certainly downplays the value of the evidence in the case. To prevent these two fallacies, it is necessary to compare the likelihood of the evidence for guilty suspects to the likelihood for innocent suspects. The likelihood ratio is the ratio of posterior odds in favor of guilt (i.e., odds after presentation of evidence)

to the prior odds in favor of guilt (odds before the presentation of evidence). Aitken argues that the likelihood ratio is the ''best'' way to evaluate the worth or value of evidence.

Aitken then presents ways to model probability when uncertainty is present. First, surveys of a relevant population must be conducted to determine the likelihood of the evidence for innocent parties. This survey could include individuals comparable to the suspect on salient factors (e.g., a sample of suspects charged in other, similar crimes). If it is known that a white male committed a crime, then the relevant population can include only white men. This will allow one to determine the ''rarity or otherwise'' of any evidence. Aitken offers a simple illustration of ''rarity'': a witness to a crime says the perpetrator had two arms. A suspect is located who has two arms. While the suspect is similar to the perpetrator on this factor, the characteristic is too common to have value. Information that a perpetrator had only one arm, however, would be valuable because such a condition is rare. Chapter four presents important techniques and formulae for computing the likelihood of a match for innocent suspects.

Chapter five discusses the likelihood ratio in greater detail. One of its strengths is the ability to logically combine the probabilities associated with several independent pieces of evidence. Assume, for example, that two evidentiary factors are present, each with a probability of 0.7. The use of standard probability theory would tend to show that combining the evidence reduces the overall value of the evidence (i.e., $0.7 \times 0.7 = .49$, which is lower than either factor's individual probability). Use of the Bayesian approach and the formulae provided, however, would yield a combined probability of .84, which strengthens the probability of guilt given additional evidence. He presents several enlightening examples to illustrate this concept and the use of the likelihood ratio as the value of any given piece of evidence. Also important is the direction of transfer. Was evidence transferred to criminals from the crime scene (e.g., broken glass fragments) or from them to the crime scene (e.g., blood stains after a struggle). Formulae for both scenarios are presented.

The true applications begin in the final three chapters, which contain formulae for discrete and continuous data, and DNA profiling. The likelihood ratios for many different scenarios are presented, varying from simple (e.g., one blood stain and one offender) to the very complex (e.g., many blood stains of many blood types and many offenders). It is in these final chapters that the book begins to have practicality. Aitken discusses the calculation and modification of formulae under a large variety of circumstances, liberally documenting the reasoning behind each computation.

Chapter six concludes with two interesting sections on calculating likelihood of paternity and likelihood of matches based on parental phenotypes when a victim's phenotype information cannot be obtained (e.g., the victim is missing). Chapter seven discusses the determination of the probability density function for continuous data using the *kernel density estimation* procedure. The final chapter briefly discusses DNA profiling and argues that while the usefulness of likelihood ratios in DNA cases has been debated, it is a very useful technique. Aitken concludes his book by presenting and defending several uses of likelihood ratios in DNA profiling.

In summary, the book is an excellent reference for the serious forensics scholar. One strength of this book is the carefully selected examples; they are simple enough to

illustrate the material for the novice statistician, yet ''real'' enough to be plausible and have meaning for experts in the field.

Jon'a Meyer
Department of Sociology
Rutgers University
311 N. Fifth Street
Camden, NJ 08102
U.S.A.
Phone: 609-225-6013
e-mail: jfmeyer@crab.rutgers.edu

Barry Edmonston (ed.). Statistics on U.S. Immigration: An Assessment of Data Needs for Future Research. National Academy Press: Washington, DC, 1996. ISBN 0-309-05275-0. 81pp. 23.95 USD.

The Committee on National Statistics and the National Research Council Committee on Population held a workshop in 1992 in order to explore ways of improving the collection and analysis of data on immigrants to better serve the needs of the federal statistical agencies and the social science community.

The current volume summarizes the discussion on immigration data needs from four of the workshops: on immigration trends, effects of immigration and assimilation, labor force issues, and family and social networks. In addition to an introductory chapter presenting the workshops' conclusions and recommendations, there is also a general chapter on data needs and another on the value of longitudinal studies of immigration. There is no section that attempts to synthesize the recommendations or relate them to any coherent set of principles informing immigration policy, perhaps because we are lacking such a unified policy perspective at the national level. The recommendations are presented piecemeal in their respective areas; whatever unity there is stems from the emphasis on more and better data and better methods for their analysis.

The workshop on immigration trends pointed to the need for more accurate figures on immigrant numbers, noting, for example, that the last decennial census may have missed anywhere from one to two million foreign born residents. This is an urgent need since the scale of immigration, both legal and illegal, is on the rise. In addition to numbers, more specific information is needed on nativity, country of birth, and dates of immigration for both respondents and their parents. The call for data on race and ethnicity could have restricted it only to ethnic origins; there are very few scientifically valid uses for the current Office of Management and Budget constructs of race which, at least in survey and census applications, produce wildly discrepant results.

The immigration and assimilation workshop looked at the effects of immigration from two perspectives: first, what happens to the immigrants as the result of their experience and, second, what effect do the immigrants have on the country. The need for better data on the adaptation of the immigrants is crucial, because members of different ethnic groups adjust differently each having its distinct problems and needs. Data on the overall effects of migration on the host country are needed to resolve the current political debates over the desirability of immigrants. In their zeal to use migrants as scapegoats, some

politicians have focused only on negative aspects of immigration, forgetting that there are many positive ones as well.

The workshop on labor force issues called for more group-specific data on labor markets noting that overly general categories such as Asian or Hispanic offer no aid to comparative analysis. Since much of the immigration unfolds in group and location-specific patterns we need separate information for the labor force effects of each group, and also for period effects at different times.

The social and family network workshop offered perhaps the most useful recommendations for improved data collection. In its recognition of the basic units of immigrant adaptation, namely the various kinship and locality-based social networks, the group called new data that considers the dynamics of the social units actually involved in the immigration dynamic, rather than relying on the larger abstractions favored by statistical agencies. By focusing on grounded variables that have proven significance in the lives of the immigrants over time, network analysis offers the greatest potential for improving immigrant data needed for sound policy decisions.

The chapter on data needs focused on a number of sources that potentially could provide new or better immigration data. Some are already being mined for data, but with refinements could produce even better information. The recommendations include adding questions on individual and parental nativity to the decennial census; local-contextual data to the Census's Public Use Microdata Sample (PUMS); more detailed immigration-related questions to the Current Population Survey; establishing joint U.S.–Mexico surveys on immigration; and a new survey of green card applicants.

The final chapter examines the usefulness of longitudinal studies of immigration and examines a number of existing longitudinal surveys that have immigrant components. The group recognizes that mounting totally new large scale immigrant surveys would be very expensive and most of the existing surveys are either restricted in their coverage of immigrant groups or in the amount or quality of data they collect. One solution recommended by the workshop was to mount smaller special-purpose surveys of selected immigrant populations that would not be as expensive as large national surveys. Another partial solution would be to modify existing surveys to collect more pertinent data, but that, of course, is often difficult to accomplish. The group also looked at alternatives to mounting new longitudinal surveys and found that some of them would be valuable for augmenting currently available information. Those worth mentioning involve oversampling groups of special interest, increasing the use of administrative records, and collecting ethnographic data for fuller understanding of immigrant adjustment. Ethnographic observations, however, would require further survey data to gauge their generality which could not be determined from the observation alone.

In attempting to evaluate the themes emerging from the workshops the reader may applaud the overall emphasis on further refining the categories involved in the collection of immigrant data, especially the focus on getting breakdowns of data at the level of ethnic/cultural units which represent the natural clusters of patterned behavior, rather than the larger artificial constructions based on aggregated data from disparate ethnic populations. The call for data on social and family networks, especially through ethnographic studies, is also an improvement over the analysis of aggregated data which conceals the actual distribution of behaviors. The emphasis on the acquisition of longitudinal

data which can reveal the dynamics and trends of the immigrant experience and the changing social forces to which they respond is also a good one, albeit difficult to achieve, at least on a large enough scale to be of real significance. The recommended stopgap measures based on the utilization of existing survey data sources may be necessary until new or improved surveys can be mounted. Even they can produce useful data if appropriate temporal and contextual data can be collected simultaneously to relate the data to the specific social ecologies of the immigrant experience.

The main problem with the presentation of the workshop findings is in the lack of any coherent framework in terms of which the reader could evaluate the priorities of data needs and directions for future research. The topics of the introductory section merely reflect the diverse interests of the workshop participants with no attempt at drawing together or synthesizing disparate elements. Still, overall, the volume presents a wealth of ideas for reforming the processes through which immigration statistics are collected, analyzed and utilized and can be recommended to anyone interested in better immigrant data.

Matt T. Salo
U.S. Bureau of the Census
Statistical Research Division
Suitland FOB #4, Room 3232
Washington, DC 20233, U.S.A.
Tel: 301-457-4992
email: Matt.T.Salo@ccmail.census.gov

G.G. Roussas. A Course in Mathematical Statistics (2nd Edition). Academic Press San Diego, CA, 1997. ISBN 0-12-599315-3. 572pp. (cloth). 59.95 USD.

This is the second edition of a book published over twenty years ago (1973) as *A First Course in Mathematical Statistics*. The author states the text is intended for use in an upper-level undergraduate course or a first-year graduate level course on mathematical statistics. To this end, a year-long lesson plan envelops most of the book's chapters. The focus of the book is to introduce the student to the theoretical constructs which underlie modern statistics; rigor is less important than explainability, but explanations are not so simple as to lose sight of the theoretical content. To this end, the student is not expected to have had an in-depth background in mathematical statistics, but is expected to have had a modicum of calculus (3 semesters) and some familiarity with linear algebra.

With most recent texts on statistics, references to particular computer programs (sometimes included with the text itself) are fairly commonplace, but not in this one. In fact, the author hopes that there is a place for a text which challenges readers to ''think'' and not delegate the thinking to a computer. His point is well taken. Sometimes statisticians and practitioners (such as myself) need to review the use of our methods for theoretical rigor. Further, students must realize early in the process of learning that the computer is a tool and that the real meat of statistics is the mathematical fabric in which it is woven. However, because the text is written as a challenge to the emphasis on computers in the world of mathematical statistics, it fails to dovetail significantly with needs of students and practitioners needs in important ways. In other words, the book could have been re-written

in such a fashion that it would explain processes often left to a computer, while challenging the reader to think about those processes and their implications.

The book consists of 20 chapters and three appendices. It also includes various formulae inside the front and back covers for easy reference. Appendix I briefly covers some topics in vector and matrix Algebra, while Appendix II includes formulae for non-central distributions. Appendix III is a series of tables covering different types of distributions (binomial, Poisson, and normal), as well as t, F and Chi-square statistics. The book also includes a short section on notation and abbreviations as well as answers to selected exercises.

Chapter 1 includes notations and definitions regarding set theory and the use of fields within the context of this text. Chapter 2 introduces the concept of probability and defines conditional probability as well as independence and combinatorial results. The author added sections that might not be taught in some introductory level classes, but would be useful at more advanced levels. In Chapter 2, for instance, the sections on product probability spacing and the probability of matchings are presented as optional topics for the introductory level student. Advanced topics are set apart throughout the text and are highlighted in the preface to the second edition.

Chapter 3 introduces the probability distribution functions of some simple random variables, namely binomial, Poisson, and normal distributions. Chapter 4 introduces the concept of probability density and links it back to the mathematical functions which defined those probability distributions. Additionally, marginal and conditional probability functions are also introduced here. Chapter 5 introduces the concepts of variance and covariance and the interpretation of the correlation coefficient. Chapter 6 discusses general properties of characteristic functions. Chapter 7 establishes the criteria for independence. Chapter 8 establishes the Central Limit Theorem and provides some immediate applications for it, as well as proofs and related applications for the laws of large numbers. Chapter 9 discusses the univariate and multivariate transformations of random variables and random vectors and applies the discussion to the t and F distributions. Chapter 10 ends the first half of the book and its focus on probability and distribution theory by focusing on order statistics and deriving some distributions used in the second half of the book which focuses on statistical inference.

Chapter 11 defines the concepts of sufficiency, completeness and unbiasedness (uniqueness). In Chapter 12, the criteria for selecting an estimator are established and applied variously to the maximum likelihood approach, the decision-theoretic approach and in finding Bayes estimators (among others). Chapter 13 develops the criteria for testing hypotheses with discussion of Type I and Type II errors, and test parameters for a normal distribution and deriving likelihood ratio statistics. Chapter 14 defines sequential sampling and derives a sequential probability ratio test. Chapter 15 defines, derives, and applies confidence and tolerance intervals. Chapter 16 introduces the general linear model and least squares estimation, as well as the derivation of the F statistic. The most practical presentation seems to be Chapter 17, which introduces analysis of variance and provides numerous and interesting applications for its use. The multivariate normal distribution is introduced in Chapter 18, along with some tests for independence and a derivation of R^2, some of which is left to the student as an exercise. Chapter 19 introduces and defines quadratic forms. Finally, Chapter 20 defines nonparametric inference and derives several nonparametric tests which are also applied.

As a textbook, the exercises are numerous and to the point, being directly related to the material covered in each preceding section. As a rule, the exercises serve as a compliment to the material, challenging the reader to 'discover' important relationships between theorems developed earlier in the text, and to discover clues to future uses of those theorems and their corollaries. For example, Exercise 18.3.7 (Chapter 18, Section 3, Exercise 7), challenges the student to discover the range of values for R^2, given the derivation of R (the sample correlation coefficient) in the text.

As one would expect with a second edition, the text seemed to be relatively free of typographical errors, and was generally well organized. The index was especially nice, providing references that a beginner might find useful and a more seasoned student of statistics would find time saving.

This book should be used to teach advanced undergraduates and/or beginning graduate students who are primarily interested in mathematical statistics. For those who are learning statistics for an applied purpose, this book concentrates too much on probability theory and not enough on the nuts and bolts of analysis, i.e., ANOVA, regression and so on. On the whole, the author succeeded in removing the computer from the statistics classroom. While this provides some obvious benefits (outlined above), this is also the greatest drawback of the book. Requiring students to ''discover'' statistical distributions by looking at tables is useful to a point, but hardly as informative a tool as ''playing'' with them in a computer. I call to the reader's attention two things in light of this criticism. First, the first edition of the text was written in the early 1970s before the advent of the personal computer or even of hand-held calculators; the second edition is well written for that time period. Second, the field of statistics has changed more since the advent of the personal computer than in the 50 years previous. The reader might recall that soon after computers were being used as calculation tools, there was a fear that the personal computer would replace the more cerebral aspects of mathematics and statistics. However, personal computers have only served to advance the field farther and faster than anyone could have dreamed.

To make best use of this text, an instructor might supplement it with some computer based exercises. For example, a simple spreadsheet that illustrates many of the exercises that were meticulously worked out in the text would serve as an excellent bridge from yesterday to today. Not all of the exercises should or could be put into a spreadsheet, but some ought to be (especially those dealing with the properties of distribution functions, probability densities and the like). On the whole, this is an excellent text, lending itself to many uses inside the classroom. However, it saddens me to recall 'how it was done' in an earlier era, and to suggest that it might not be useful today. The rigor and attention to detail that went into this work does make it useful today, even without a computer.

John Booher
Center for Public Leadership Studies
George Bush School of Government and Public Service
Texas A&M University
College Station, TX 77843, U.S.A.
Tel: 512-349-6641
email: John_Booher@exchange.tcada.state.tx.us

Minitab, Release 11 for Windows. Minitab Inc., U.K. 1996.

Programming without programmers with Minitab

Minitab and other software packages at a statistical agency

Within the university sector, Minitab is a well-known statistical software, with a reputation for being both teacher and student friendly. While teaching at the university, we started to use Minitab in 1978 and, due to the straightforward commands in Minitab, we found it possible to discuss concepts and understanding with the students instead of wasting time on computer technicalities.

At a statistical agency, however, the users and applications are different. We will here discuss different kinds of software used for statistical purposes at a statistical agency and whether Minitab can be a suitable tool in that environment.

At Statistics Sweden we are now leaving all mainframe applications and future applications will be in a PC/net environment where all data sets are stored in SQL databases. We want to distinguish between the following tasks, where we are of the opinion that Minitab can be an alternative software for tasks 2, 3, and 4.

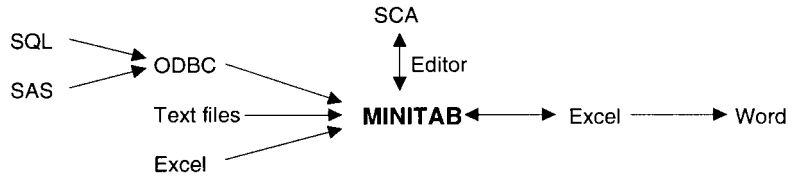
Different statistical tasks	**Software generally used today**
1. Processing of large registers	SQL, SAS, SuperCross
2. Processing of data from sample surveys	SQL, SAS, SPSS
3. Processing of time series	The X11 procedure in SAS/ETS
4. Data analysis	EXCEL, SAS, SPSS
5. Preparation of statistical publications	EXCEL, WORD

The production at a statistical agency requires quality assurance – all steps during the statistical processing must be documented thus enabling detection of errors and minimizing the risk of error occurrence. It is of vital importance that this documentation is easily understood – if only the programmers can search for errors the risk of errors being undetected will increase. In many cases the same kind of processing is repeated every month, quarter, or year. Command-based software is therefore required for the tasks 1–4 above.

Minitab can be used either via menus or via commands but the commands will always be stored in a History Window if you want to check or document what you have done. You can also use Minitab Macros stored in text files to automate repetitive tasks. The macro language consists of straightforward and powerful commands and when you develop your own macros you have full documentation and control of your statistical processing. We have noticed that Excel is often used for statistical computing – it seems easy to do the computing with Excel but how do you check and document your computations?

Metadata are important and the statistical software should make it easy to save metadata together with the statistical data set. We have found that software, where data are stored in easily accessible worksheets where you can mix text and statistical data, can improve the collection and documentation of metadata. To look at data in the worksheet while you execute commands which process data is also an excellent method to discover errors. Minitab works well in all these respects.

At a statistical agency we have access to many kinds of software. It is therefore not important that one single package can do everything – the software we use must instead be able to interact with other software. The new version of Minitab functions well in this respect – you can import data in many ways and communicate with other software, e.g., SCA or X12 for advanced time series analysis, and export results for presentation with Excel. For us, Minitab has the central role in the processing, supported by other software:



Release 11 for Windows – important features and news

Minitab has a complete menu interface which gives good support to new users. Advanced users may prefer the interactive command-line option. On-line Help exists for all dialogue boxes and all session commands. The documentation consists of a User's Guide and Reference Manual which are well written.

The macro programming language is perhaps one of the most appealing features of Minitab. The less experienced user can make simple macros with ordinary session commands. If you are more advanced you can combine session commands with DO loops and conditional statements and build powerful macros suitable for your tasks and your way of working with data.

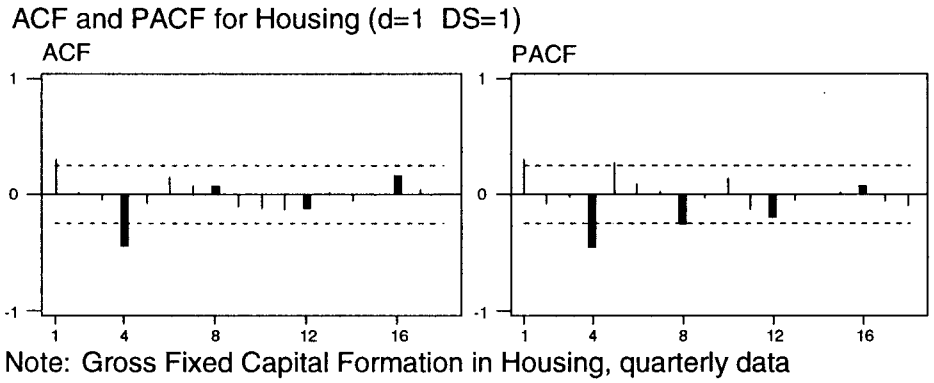
Release 11 can be used on Windows 3.1, 3.11, Windows 95 or Windows NT. The size of the worksheet allows 1,000 columns with numeric or alphanumeric data, 1,000 constants and 100 matrixes. The total amount of data is limited only by the amount of available memory – our (old) PC allows 15,000,000 cells with data in double precision. As many data sets at Statistics Sweden are stored in SQL data bases the new way of querying data bases using ODBC is of great value. Release 11 is well adapted to Windows 95. Long file and variable names are allowed, right click gives pop-up menus and communication with other software compatible with Windows 95 is easy via the Clipboard or via the DDE option which automatically transfers data.

For many analyses, graphs are produced automatically. Regression analysis, e.g., includes graphical residual analysis in four charts. You can also produce graphs with simple menus or more advanced session commands. Graph editing and brushing are easy tasks. Logistic regression is a new feature in Release 11 – many multivariate methods are now included in Minitab.

How can you use Minitab?

1. Create your own commands. You can tailor Minitab to suit your way of working with data. E.g., if you work with time series analysis you can create a number of macros which

support your way of identifying and checking time series models. Our macro for identifying ARIMA models gives the following result:



With our macro we get four charts (d/DS = 0/1) to identify the ARIMA model for a seasonal series. We want acf and pacf side by side and the seasonal correlations marked. Our minitab command is:

```
MTB > %iden 'Housing';
SUBC > s 4.
```

2. Standardize monthly processing. If you produce monthly statistics you can check the results for last month by analyzing the standardized ARIMA residuals for all series before publishing. To develop a simple Minitab macro is an effective way of doing these kinds of quality control.

3. Sizable isolated tasks. We have recalculated 72 monthly labor force surveys with about 15,000 observations each month. A Minitab macro was produced which imported each survey and calculated and saved new sample weights for estimation with the new calibration (Deville and Särndal 1992). The calibration used 106 constraints.

Conclusion

Can Minitab be a suitable tool at a statistical agency? We think that Minitab, or a similar statistical package, should replace statistical computing with Excel. Excel is easy to use but can easily be a source of miscalculations. Programmers will always be rare persons – if you want to be able to build your own statistical applications we think that Minitab can be a good alternative. We think that the problem with Minitab is its name – and this problem will go from bad to worse as the capacity of Minitab becomes better and better. The name may give the impression that Minitab is suitable for simple applications only. That is not the case nowadays.

Contacting Minitab

Tel: 1-814-238-3280 (U.S.) or +44-(0) 1203-695730 (U.K.)
Fax: 1-814-238-4383 (U.S.) or +44-(0) 1203-695731 (U.K.)
E-mail: sales@minitab.com (U.S.) or sales@minitab.co.uk (U.K.)
URL: http://www.minitab.com

References

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Britt Wallgren and Anders Wallgren
Statistics Sweden
Department for Research and Development
SE-701 89 Örebro, Sweden