Journal of Official Statistics, Vol. 23, No. 4, 2007, pp. 593–602

Book and Software Reviews

Books for review are to be sent to the Book Review Editor Jaki S. McCarthy, USDA/NASS, Research and Development Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.
Email: jaki_mccarthy@nass.usda.gov

David J. Hand. *Information Generation: How Data Rule Our World.* Oxford: Oneworld Publications 2007, ISBN 1-85168-445-X, 246 pp, 25.95 USD.

I have two favorite chapters in David Hand's *Information Generation: How Data Rule our World*. The fourth chapter entitled "Big Brother's Eyes" is an excellent discussion of the tensions posed by data: privacy versus efficiency and efficacy. His compelling and thought-provoking examples range from RFID and bar codes to surveillance cameras and crime data.

Even readers who have given serious thought to the changing norms of privacy in our society may learn from some of Hand's insights. For example, he nicely frames privacy questions as a balance between personal privacy and public goods. "While collecting detailed data on individuals may threaten their individual privacy, it is very clear that it has the potential for immense good for the group." He continues with examples from epidemiology, post-marketing surveillance of pharmaceuticals, anti-terrorism efforts, and fraud detection by banks. A few pages later, Hand describes the increasing segmentation seen in political campaigns. He writes, "It used to be the case that what mattered was what the voters knew about the candidates, so they could decide who to vote for, but this has changed. What matters now is what the candidate knows about the voters. With this knowledge, the candidate can decide who to speak to and what to tell them."

The other chapter I liked immensely is the sixth, "In Data We Trust," which is an exposition on data quality. Hand articulates concepts to characterize quality such as accuracy, precision and bias. Under "distorted data," he explicates measurement error of various sorts. In missing data and detecting errors in data, the reader learns that data are inevitably flawed, very rarely for malicious or negligent reasons. The chapter also includes helpful suggestions on how to detect errors and live with them.

A final section of this chapter is called "The morality of good data." Here, Hand cautions against overly exalting good data (that is, data not exhibiting the various flaws discussed throughout the rest of the chapter). [Good data], he writes, "can lead to correct conclusions. That they need not do so is because data, in themselves, do not tell us anything: data have to

be analyzed, have to have their meaning squeezed from them, in order to tell us things, and it is entirely possible that mistakes may be made in the analysis" (p. 205). This section is spot on, but I wish that he had gone further with one other way in which good data does not necessary lead to good information: with floods of data comes the problem of knowing what to ask. Whether FBI files prior to September 11, 2001, or drug trials data indicating risks of a medicine, ex-post analysis often clearly shows a smoking gun. But regular users of data know that choosing the right questions is a difficult challenge. That having too much data can obscure as much as too little data is a difficult lesson to learn for many inexperienced data users. Even without making this additional point, this section nicely complements an already informative discussion of "bad" versus "good" data.

These two chapters have some things in common. They remind/inform lay and expert readers how the intricacies of data and information infiltrate our lives in ways that we may and many that we may not always realize. The chapters are filled with real-life examples that are illuminated using technical concepts. Many readers will probably return to these ideas in the days after first reading about them.

The rest of the book is also wide-ranging and conversational in tone. Hand does an admirable job covering the many dimensions of data and data usage in our society. Perhaps because of the ambitious sweep of the book, there can be unevenness to the delivery. Some chapters are very friendly to novice readers, while others tackle sufficiently complex material that only very diligent readers will be able to grasp the concepts introduced.

The first chapter, "Let there be light," introduces the concepts of data, theory and information. Throughout, Hand points to both the tremendous power of data and to the uselessness of data used poorly. One lesson might be: we are awash with data, overwhelmed by the costs of collection, storage and analysis. Data can be powerful, but without valid interpretation, data are just noise. Hand writes, "Information, then, is *the useful content of data*, the extent to which the data permit us to say something helpful about the phenomena we wish to understand" (p. 10).

In fourteen pages, Chapter Two traces the invention of numbers, the beginning of counting and measuring, and initial recording systems forward to the unfathomable numbers of data points in our biggest data sets today. The focus on numbers can understate the increasing importance of nonnumeric (especially geographic) data.

Chapter Three is a longer chapter that lays out the evolution of early science, the scientific method, and the "mathematization of science." The story is filled with delightful and vivid anecdotes of such key figures in the history of science as Galileo and Newton. Hand shows that pettiness and basic human weaknesses are often a part of the scientific process. He also argues repeatedly that a strength of science is its fallibility and its ability to self-correct. Models, description, explanation and prediction are all discussed and contrasted, although perhaps not fully distinguished. The last four pages of this chapter address the relationship between science and religion. Hand does differentiate the two, but he does not tap into or illuminate any of the tension and fury of the current debate.

Modern data science, we learn in Chapter Five, is about "how to extract information from data" (p. 156). This meaty chapter sketches out principles of data design, outlines basic concepts in statistics, and introduces other computer-age data sciences such as database technology and machine learning. The chapter also alludes to data compression and data display.

In Chapter Seven, Hand presents Charles Babbage's taxonomy of dishonesty and deception with data: hoaxing, forging, trimming, cooking. Fortunately, this chapter is brief, otherwise the hair-raising stories of bad behavior would lead many readers to despair.

Throughout the book, Hand includes many classic data stories: O-rings in the Challenger space shuttle, Literary Digest pronouncing Dewey over Truman, the Lancashire milk study. His chatty and opinionated style makes for easy and pleasurable reading.

Just as I found some favorite chapters, other chapters felt as if they targeted a different reader. However, almost anyone interested in data and information and how they penetrate our society will find something entertaining and informative in this book.

A. Rupa Datta
NORC at the University of Chicago
2140 Shattuck Avenue, Suite 307
Berkeley CA 94704
U.S.A.
Phone: 510-647-3660
Fax: 510-647-3661
Email: datta-rupa@norc.uchicago.edu

Peter Olofsson (2006). *Probabilities: The Little Numbers that Rule Our Lives*. Hoboken, NJ:Wiley, ISBN 0470040017, 262 pp, $59.95.

> In a society that has packaged basic knowledge in pill form, a student, needing some learning, goes to the pharmacy and asks what kind of knowledge pills are available. The pharmacist says "Here's a pill for English literature." The student takes the pill and swallows it and has new knowledge about English literature!
>
> "What else do you have?" asks the student. "Well, I have pills for art history, biology, and world history," replies the pharmacist. The student asks "Do you have a pill for statistics?" The pharmacist brings out a huge pill twice the size of a jawbreaker and sets it on the counter. "I have to take that huge pill for statistics?" inquires the student. The pharmacist replies "Well, you know statistics always was a little hard to swallow."

If you enjoy puzzling over hard-to-swallow statistical curiosities you will find pleasure in "Probabilities: The Little Numbers that Rule Our Lives." This book offers a guided tour through the counter-intuitive mind-benders of statistics. The result is a Readers Digest approach to "popular" statistics, a series of vignettes and stories, often with suggested uses for party demonstrations.

Peter Olofsson is a professor of Mathematics at Tulane University in New Orleans. This, Olofsson's second book, is written for general audiences with a taste for puzzles and at least an elementary statistics course under their belts. The book was written throughout Hurricanes Katrina and Rita and related evacuations to Houston, West Texas, and New Mexico. The presence of the hurricanes and refugee experiences are felt throughout examples used in the book. In examining the tendency of rare occurrences to occur in clumps, he notes that in Houston, many residents offered that they were "due" for a big one, since no large hurricane had recently hit, while in New Orleans, residents were

placated by the knowledge they had not been hit by a big one in quite some time. He further illustrates with the macabre example of a doctor who tells you that you have a fifty–fifty chance of survival and adds, "You're lucky – my last patient died."

You've probably heard the one about the two statisticians arguing in a bar. . . .

The book begins with a basic introduction to probability, probabilists, conditionals and combinatorics. Throughout this introductory chapter the theme of the book is evident: anecdotal stories illustrate the concepts, followed by the mathematical formula. Most of the stories are humorous, some historical, a few tragic. The illustration of principles of probability theory by way of real-life situations provides a valuable resource for anyone teaching statistics who wants some sly and saucy examples to use. Professor Olofsson is undoubtedly a delight to his classes. The wise reader will use the social recommendations with restraint, however, as the conversation starters and party games designed to help the reader "Win Money and Lose Friends" are likely to be equally effective.

If I've told you n times, I've told you n + 1. . .

"Probabilities" is not a text and works best as a refresher to the rusty. Those with a freshly comprehensive grasp of the math underlying probability theory will find themselves "browsing" for stories, while those without a basic understanding of probability will find themselves lost. Fortunately for all, the stories abound and are the foundation of the narrative. With the basic language of probability established early on, the book moves quickly through tiny, backward and inevitable probabilities.

Tiny probabilities are the seemingly impossible situations that one encounters many times along life's meandering paths. With each breath you take, for example, what are the chances that you just inhaled a molecule that was exhaled by the dying Julius Caesar in his last breath in 44 B.C.? You may remember this calculation from the original work of Sir James Jeans's 1940 book An Introduction to the Kinetic Theory of Gases, or you may have encountered it elsewhere and know that your chances are better than 50–50. Olofsson continues with another example you are less likely to have heard, of the time in 1987 when he went for a swim on the east coast of Australia and lost his glasses. In 1992 he encountered a statistician from Brisbane, near the very spot where Olofsson lost his glasses. Although the man had not seen the lost glasses, Olofsson retells the story as an illustration that one remarkable coincidence, despite its very low specific probability, exists in a sea of possibilities, and the chance of any one particular coincidence is actually quite unremarkable.

If there's a 50–50 chance something will go wrong, nine times out of ten, it will.

Backward probabilities rely on Bayes Theorem, which were published posthumously, to debunk the assertions that 75% of all car accidents are caused by sober people (so the drunkard should always drive) and 100% of all divorces begin with a marriage (so marriage is a bad idea). Probabilities are calculated backward from an observed outcome frequently in courts, where statistical expert witnesses testify to the odds of a particular event in order to demonstrate a reasonable doubt, or lack thereof. Olofsson explores the case of Sally Clark, convicted of murdering her two children after each died shortly after birth, about one year apart. An expert witness, a pediatrician rather than a statistician, claimed that there was a 1 in 73 million chance of having two such deaths in one family, and Clark

was convicted on this basis. Since the deaths were not independent (there are genetic factors to consider), the odds were later recalculated using Bayes' rule to about one in 100. Olofsson reworks the odds again, this time weighing against the probability of another rare event – double infanticide – arriving at a fifty–fifty probability of Sally's innocence. (Sally Clark was released in 2003 after serving three years in prison, and she herself died tragically in 2006.)

I asked a statistician for her phone number. . . and she gave me an estimate.

The irony of an expected value is that it is so often not expected, or even possible as an individual outcome. No one really expects to have 2.2 children, for example, but it is often cited as descriptive of the "typical" family. In a discussion that ranges from gambling and the stock market to testing the blood of American draftees (in pooled samples), Olofsson looks for the unexpected to exemplify expected results. The waiting time paradox is used, for example, to calculate how long an alien visiting the earth will wait to experience a great (magnitude 8 or higher) earthquake. These earthquakes occur an average of once per year. At the time an earthquake occurs the expected waiting time will be one year, and the expected time since the previous quake is also one year. An intuitive but statistically naïve alien might expect to wait about a year upon landing for the next earthquake to occur. The paradox is that an average interval of one year may be the result of a two-year interval and a one-day interval. The alien is far more likely to land during the two year interval than the one-day interval, however, making the average waiting time greater than the average interval. If you are trying to catch a bus or visit Old Faithful (which erupts every 90 minutes on average), you can calculate an expected wait time and propitious arrival.

> Three statisticians went hunting and came across a large deer. The first statistician fired and missed, by a meter to the left. The second fired and missed, by a meter to the right. The third didn't fire at all, but shouted in triumph, "On the average we got it!"

One of the delights of this book is Olofsson's inclination to seek out intuitive, but false implications of probability theory and to demonstrate the fallacy behind them. Thus, the reader is reminded that although no one escapes the law of averages, the probability of an equal number of heads and tails gets smaller and smaller the more times you toss it. Like gambling, the purchase of insurance represents an expected loss. Perceived randomness is often found in the uncertainty of early conditions, so that a coin is slightly more likely to come up the way it started than the other way. Prognosticators of probability at all levels are likely to find interesting and relevant illustrations in this book that they will use again and again.

Amy Flowers
Market Decisions
75 Washington Ave.
Portland, ME 04210
U.S.A.
Phone: 207-782-3977
Fax:207-767-8158
Email: flowers.amy@gmail.com

Sergey Dorofeev and Peter Grant (2006). *Statistics for Real-Life Sample Surveys: Non-Simple-Random Samples and Weighted Data*. Cambridge University Press, ISBN 0-521-67465-4 (pbk), ix, 266 pp, 45.00 USD.

Delivering impressively on the promise of its title and subtitle, this book provides a very useful compendium of statistical concepts, tools, and techniques for dealing with the realities of survey sampling in actual practice. Dorofeev and Grant, both from the Australian polling firm of Roy Morgan Research, have drawn on their many years of practical experience to bring together key materials on sampling and weighting, and on associated issues in significance testing and multivariate analysis. That applied experience is also the source of numerous real-data examples that effectively illustrate the methods.

Written in a clear, lean, and engaging style, the book covers its selected topics systematically and thoroughly. Mathematical details are provided in appendices. The book does not attempt to be encyclopedic, but it would serve well as a reference for the practicing survey researcher or as a supplement to textbook treatments of sampling or of survey methods in general.

As the authors note, most of what they cover could be found elsewhere in the published literature upon a diligent search. However, besides being widely scattered, the relevant materials often miss the mark for survey research as it is actually practiced. Many such sources focus on simple random sampling, whereas in practice nonsimple (and/or nonrandom) samples are the rule rather than the exception. Other sources address complex probability samples, but often at a mathematical level that presents practical barriers for the typical real-world survey researcher.

Dorofeev and Grant, by contrast, provide detailed guidance to improve the survey practitioner's handling of stratified samples, cluster samples, multi-stage samples, and other departures from simple random sampling. Appropriately, they devote most of their attention to nonsimple samples that are, however, probability-based. Nonrandom sampling is covered adequately, but not extensively. That lesser emphasis follows from the fact that statistics as such, and especially inferential statistics, are of limited value for quota, convenience, purposive or other nonprobability samples.

The book is organized into five chapters and eight appendices, plus references and an index. Appendix A is a concise and admirably precise review of the central concepts and terms in elementary statistics. The information therein is densely packed, but sufficient in scope and depth to make the rest of the book accessible to those survey practitioners whose statistical training is either shaky or rusty. Chapter 1, which covers a wide variety of probability-based and nonprobability sampling methods, is pitched at only a modestly higher level. As such, it serves to review and define the concepts and terms specific to sampling, and to set the stage for the two core chapters that follow.

The heart of the book lies in Chapter 2, on the mechanics of weighting, and in Chapter 3, on the consequences of weighting for inferential statistics. Weighting of survey data is required with disproportionally stratified samples and is also widely used in post-stratification, to bring the distribution of survey respondents into line with known characteristics of the population. Chapter 2 details the calculation of weights by

several common methods, such as cell weighting, marginal weighting, iterative raking, and hybrid weighting systems. The chapter concludes with a brief but important discussion about the ethics of full disclosure in research reports that are based on weighted data.

While weighting may reduce bias in the estimation of parameters, it will also increase sampling error in those very estimates. Chapter 3, the most mathematical of the five chapters, deals with the calculations required to adjust the variance of estimated parameters to take into account not only sampling complexity (like stratification or clustering) but also the effects of weighting. To account rigorously for all aspects of the sample design, a unique variance calculation should be performed for each separate parameter being estimated. In practice, such calculations are rarely undertaken. Dorofeev and Grant suggest what they call the "calibrated sample size" as an approximate but more practical tool for incorporating the overall effects of sampling and weighting when conducting statistical inference on survey data. Survey practitioners are likely to find this suggestion quite useful.

Chapter 3 next offers a measure of the utility of weighting based on the ratio of (a) the difference between unweighted and weighted parameter estimates (as an indication of the likely reduction in bias) to (b) the difference between the raw and "calibrated" standard errors. The authors then provide some cautionary notes about how statistical software may mishandle weights, with disastrous consequences for analysis if the mishandling goes unnoticed by the analyst. The chapter concludes with imputation for missing data and its effects on the variance of estimators.

Chapter 4 addresses the uses and abuses of significance testing, especially with reference to tests of means or proportions. The chi-square and Kolmogorov-Smirnov tests are also covered. While it complements the solid review of basic statistical concepts provided in Appendix A and Chapter 1, this chapter is a bit disappointing in that it does not focus to any great degree on issues specific to sampling and weighting. That slight disappointment also extends to Chapter 5, where the discussion moves from analysis of contingency tables to regression analysis, cluster analysis, and data fusion. Again, there is relatively little in this chapter that is specific to issues in sampling and/or weighting. Nor is it obvious that the analytic methods chosen for review in the chapter are necessarily the key ones for applied survey research. Survey researchers will likely find the main contributions of the book in Chapters 2 and 3.

Four of the appendices provide proofs to support the results discussed less mathematically in various chapters. Another appendix includes tables of statistical distributions (such as the standard normal) that are also readily available in many other sources. In just two pages, Appendix B offers a few suggestions for further reading and some citations to the list of references. The reference list is also rather brief, and only five of the 61 sources in the list were published after 1999. The index seems nicely detailed and accurate. Indeed, the book as a whole is almost entirely free of typographical or substantive errors.

In sum, Dorofeev and Grant have provided the practicing survey researcher with a most valuable reference and guidebook to issues of sampling and weighting with nonsimple and/or nonrandom samples.

Burke D. Grandjean
University of Wyoming
Departments of Statistics and Sociology
and Wyoming Survey & Analysis Center
Dept. 3925
1000 E. University Avenue
Laramie, WY 82071
U.S.A
Phone: 307-760-5913
Fax: 307-766-2314
Email: burke@uwyo.edu

Wayne A. Fuller. *Measurement Error Models*. Hoboken, N.J.: John Wiley and Sons, 1987, 2006, ISBN-13 978-0-470-09571-3, 440 pp, $89.95.

Suppose you were trying to estimate the average daily caloric intake in a population based on a sample of individuals reporting what they ate on the previous day. Such an enterprise is fraught with potential errors. Among other things, 24-hour recall of food consumption, both varieties and amounts, is an extremely difficult task to ask a randomly sampled individual to perform. In addition, translating those personal intakes into calories with complete accuracy is beyond the scope of our scientific ability.

Such potential sources of error are well appreciated by survey statisticians. They are not, however, really the measurement errors meant in Wiley's paperback reissue of the classic text, *Measurement Error Models*. What the author, Wayne A. Fuller, has in mind in this context is the random variation in the calories consumed by an individual on a particular day from what he or she ingests on an average day. Variations of this sort can all but be ignored by a survey statistician interested in estimating the mean daily caloric intake in a population *per se*, as these "errors" average out in a well-designed survey. They are, however, a complicating factor when one wants to model something like the contribution of an individual's daily caloric intake to his or her blood pressure.

Developing methods for dealing with errors in variables in a modeling context and championing the use of those methods has been one of the many lynchpins of Professor Fuller's remarkable career. After attaining a PhD in agricultural economics from Iowa State University in 1959, Fuller became a faculty member of both the economics and statistics departments at ISU, where he is now a Distinguished Professor *Emeritus*. In addition to his invaluable work on the error-in-variables problem, Professor Fuller has made important contributions to time-series analysis and to survey sampling. His very first article on survey sampling (Fuller 1975) developed the linearization methodology still used in major software packages for estimating linear regression coefficients with complex survey data. It surprised me to discover that a good part of that celebrated article concerns estimation when there are errors in the explanatory variables. This may seem like a secondary issue to survey statisticians, but, as Professor Fuller has frequently insisted, it is a very important one.

Measurement Error Models is a solidly built four-story edifice. Chapter 1 (the ground floor) discusses the single-explanatory-variable linear model assuming that the explanatory variable is measured with error in a well-defined way. Chapter 2 extends the analysis to a linear model with a vector of explanatory variables. Chapter 3, the longest

chapter, allows the measurement-error variance to be nonnormal and to vary across units, perhaps even to depend on the variable values. The chapter goes on to address many types of nonlinear models. Finally, in Chapter 4, not only can the model be nonlinear and nonnormal, but the dependent variable can be a vector of values.

There are many real-world examples in *Measurement Error Models*, reflecting the broad range of applications for proper errors-in-variables modeling gleaned from a host of different disciplines. The prose is clear and precise. Professor Fuller's deep understanding of the interplay among theory, methods, and data is evident on every page and may even be slightly contagious. Nevertheless, the level of rigor will overwhelm most readers without a solid background and interest in mathematical statistics.

Each chapter is divided into sections and subsections, with exercises after each section. There are also exercises after the four technical appendices to Chapter 1 and the three to Chapter 4. *Measurement Error Models* was meant to be a graduate-level textbook as well as a single source for the up-until-that-point scattered literature on the error-in-variables problem.

Although masterful, the book is not perfect. I could find no discussion on the impact and treatment of variables measured with systematic bias, a topic of keen interest to many survey statisticians. In our daily caloric-intake example, such a systematic bias would occur if many individuals tend to underreport their actual 24-hour intakes. There is much evidence that this is the case.

Even taken on its own terms, *Measurement Error Models* suffers from being the re-release of a 1987 text, not a revised edition. Thus, the tremendous impact of McCullagh and Nelder's *Generalized Linear Models* (1989) on the way statisticians handle many types of nonlinear and nonnormal models is missing. Similarly, allusions to available software, like ISU's own EV CARP, are skimpy and mostly to be found in the preface.

A reader wanting a single volume on the errors-in-variables problem and its treatment might do better with Carroll et al. (2006). Still, as those authors' frequent references to *Measurement Error Models* attest, Professor Fuller's book remains indispensable to any serious student of the problem.

Another relevant topic is understandably missing from *Measurement Error Models* given when it was developed. Nusser et al. (1996) show how to estimate the distribution across a population of average daily intakes for a dietary-component, like calories, assuming no systematic reporting biases in an individual's first day of intakes. I had a very small part in the development of this methodology. My principal role was to assure the powers that be at the U.S. Department of Agriculture and other skeptical government agencies that the team from ISU, headed by Professor Fuller himself, was creating something of enormous value.

An updated version of *Measurement Error Models* would have been preferable. Nevertheless, a less-expensive paperback edition of the original will be a most welcome addition to many bookshelves.

Phillip S. Kott
United States Department of Agriculture
National Agricultural Statistics Service
3251 Old Lee Highway
Fairfax VA 22050
U.S.A.
E-mail: pkott@nass.usda.gov

References

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective (Second edition). Boca Raton: Chapman and Hall.

Fuller, W.A. (1975). Regression Analysis for Sample Survey. Sankhyā: The Indian Journal of Statistics, Series C, 37, 117–132.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. London: Chapman and Hall.

Nusser, S.M., Carriquiry, A.L., Dodd, K.W., and Fuller, W.A. (1996). A Semi-parametric Transformation Approach to Estimating Usual Nutrient Intake Distributions. Journal of the American Statistical Association, 91, 1440–1449.