VERSITA

Journal of Official Statistics, Vol. 29, No. 2, 2013, pp. 323–327, DOI: 10.2478/jos-2013-0025

Book 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

Gerald J. Hahn and Necip Doganaksoy. *A Career in Statistics: Beyond the Numbers*. Hoboken, NJ: John Wiley & Sons, Inc., 2011. ISBN 978-0-470-40441-6, 340 pp, $69.95.

Hahn and Doganaksoy provide a valuable service to the statistics community of academics and practitioners through writing and publishing *A Career in Statistics: Beyond the Numbers*. Their contributions for readers are many, as they provide both an overview of the field of statistics as well as specific advice for students and practitioners.

The first four chapters provide background for careers in statistics (Chapter 1) and review what statisticians do in business and industry (Chapter 2), in official government roles (Chapter 3) and in other areas of application (Chapter 4).

Chapter 5 – The Work Environment and On-the-job Challenges – is a "must-read" for any student of statistics or for any professional who needs to either learn or be reminded that most professionals that they will work with are not themselves statisticians, nor do they have training in statistics. For example, many co-workers may not understand your role as a statistician, the appropriate use of statistics for different applied problems or how statistical analysis could add value to their tasks and projects. A statistician may spend a significant amount of his or her time "marketing" his or her skills within their own organization. Subsection 5.4 addresses the issue of role delineation: Is a statistician a consultant or a member of a team? When do the roles fuse and/or change? All of this information is extremely relevant for statisticians who want to be useful and effective within their organizations and beyond.

Chapter 6 focuses on traits and behaviors of successful statisticians, with a focus on "soft" or "people" skills and an assumption that readers possess the necessary technical skills (which are however insufficient on their own) to do their jobs well. These topics may typically receive little attention in a graduate or professional skills training situation focused on statistics and numerical analysis; however, having these skills is important to the success of any professional and, the authors argue, especially professional statisticians.

Professional training and advanced degrees are the topic of Chapter 7, with the authors providing insight into graduate programs as well as the value of informal educational

experiences, such as internships and participating in consulting arrangements. Chapter 14 gives attention to lifelong learning for statisticians.

Chapter 8 focuses on the job search and the recruiting process specific to statisticians while Chapters 12 and 13 discuss different career paths, including academia.

Best considered as an on-the-job primer for statisticians, Chapters 9, 10 and 11 discuss many practical topics that can be encountered by one new to the field, or one seasoned in the field. These include: project selection, estimating project costs, and successfully executing projects. Subsection 9.4 includes key advice that seasoned practitioners would share with aspiring statisticians planning on or embarking on a career. Much of the advice reminds readers to be relevant, keep it simple and find ways to be of value to their organizations.

This book is extremely well done. The sidebars and "major takeaways" offered in the text are very useful and present quick and easy summaries for the reader. I would recommend this book to any person considering an analytical support or analytical leadership position in statistics (and even related fields). Portions of this book, if not the entire text, would be appropriate required reading for professional training for graduate students in statistics.

Heather H. Boyd, Ph.D., C.P.M.
University of Notre Dame,
Research Development Program Director
Office of the Vice President for Research
940 Grace Hall
Notre Dame, Indiana 46556
(phone) 574 631 4104
E-mail: hboyd@nd.edu

Ton de Waal, Jeroen Pannekoek, Sander Scholtus. *Handbook of Statistical Data Editing and Imputation*. New York: Wiley, 2011, ISBN 978-0-470-54280-4, Hardcover $149.95.

The handbook compiled by Ton de Waal, Jeroen Pannekoek, and Sander Scholtus of Statistics Netherlands is an enjoyable, informative, instructive, and comprehensive compendium of known methods for the editing and imputation of major surveys. Expert technical knowledge is expressed clearly on topics of statistical science, mathematics, and linear programming, with separate discussions for automated processing and interactive processing of edits. Methods for automation of editing and imputation are a focus. Tools supporting the Fellegi-Holt paradigm are emphasized (Fellegi and Holt 1976). Tools for interactive edit and manual imputation such as Blaise (Statistics Netherlands) are discussed briefly in the context of selective editing. Discussion of donor imputation using nearest neighbor imputation methodology (NIM, Bankier 1999) includes a detailed comparison with Fellegi-Holt methodology. Discussion of methods for variance estimation compares the bootstrap and jackknife methods with multiple imputation and fractional imputation. The handbook develops handling of edits through a series of mathematical theorems with proofs and clear examples of their application. The

theoretical development leads incrementally to sophisticated tools for automated edit and imputation of categorical variables as well as continuous variables. Computational methods such as *branch-and-bound* algorithms are thoroughly discussed throughout the book. Each chapter ends with a generous listing of international references. The subject index is thorough and reliable.

Automation of editing and imputation emphasizes the need for classification of edits. The authors address this early in the handbook, favoring a distinction between *hard edits* which are logically necessary for a record to be consistent (e.g., NET PAY = GROSS PAY – DEDUCTIONS) and *soft edits* which serve as guidelines to flag potential errors in the record (e.g., AGE $\leq$ 110 years). The authors define subclasses by mathematical form of the edit. Classification of edits is conceptually important for any survey group that is contemplating the use of a centralized edit repository (i.e., database) which is maintained independently of other production systems (e.g., automated edit and imputation systems). The classification of an edit might be used to determine the scope for its application. For example, the use of an edit as a prescriptive relationship amongst survey variables (e.g., adjustment procedures in Chapter 10) may be suitable for hard edits and less desirable for soft edits. The text offers best practices (e.g., strategy to mitigate overediting) and characterizes the intended use for automated procedures. The philosophy for editing and imputation adopted by the authors might not generalize to survey organizations lacking a national registry or centralized data collection. However, the technical understanding of procedures portrayed by the authors would inform any usage of the procedures.

The handbook serves as a guide suggesting options for handling of edit constraints in tandem with automated imputation of survey reports (records). The authors discuss procedures to meet three steps: 1) the edit; 2) the imputation; and 3) making the imputed values consistent with the edits. For step (2), deterministic imputation procedures and stochastic imputation procedures are described. For step (3), incorporation of edit rules into the automated construction of imputed values is discussed in Chapter 9; and adjustment of imputed values to meet edit rules is discussed in Chapter 10. Nearest neighbor imputation methodology (NIM) is discussed in the context of detection and correction of errors (e.g., Johanson 2012).

At the Washington Statistical Society's 2011 Morris Hansen Lecture hosted by the National Agricultural Statistics Service (NASS), Roderick Little spoke of the lack of congruence between the theory for modeling and the theory for sampling as a "schizophrenia." The handbook includes a useful theoretical development of the relationship between model-based imputation and sample-based weights in Subsection 7.3.4 *Connection between Imputation and Weighting*. In particular, there is the question of how to manage design weights and adjustment factors in the context of model building for imputation and estimation. The handbook partially addresses the issue with examples and theoretical proofs suggesting appropriate procedures for incorporating design weights into a model-based imputation. In particular, the authors have proven conditions (Theorem 7.1) under which estimators are equivalent across 1) weighting the respondent data by applying the *regression estimator*; 2) imputing the nonrespondents using regression imputation and then weighting the entire sample by applying the regression estimator; and 3) imputing nonrespondents and nonsampled elements using

regression imputation. The authors refer to these three cases as the *weighting approach*, the *combined approach*, and the *mass imputation approach* respectively. Further discussion of the design/model compromise (DMC) is provided by Roderick Little (Little 2012).

Being somewhat new to surveys, I considered the book in terms of my current projects: implementing a new system for significance editing called SignEdit (Kosler 2012); utilizing procedures from the Banff System commercialized by Statistics Canada (Johanson 2012); implementing iterative sequential regression (ISR) with edit constraints for the *Agricultural Resource Management Survey* (Robbins et al. 2012); and construction of a centralized edit repository. In a supportive manner, the handbook provided useful and in-depth technical background on most editing and imputation topics pertinent to these projects (e.g., donor and ISR imputation procedures). Several topics were treated with a history of the development of known methods (e.g., error localization) and a comparison of approaches (e.g., adjustment of imputed values to meet edit constraints).

The handbook's authors synthesized a broad range of material for the practicing survey statistician, gathering topics as diverse as *group random hot deck imputation*, *Gibbs sampling*, and *Fourier-Motzkin elimination*. For useful methods, the reader would find a convenient combination of textbook level descriptions of methodology, examples commonly published in hard-to-find technical reports (e.g., selective editing for the Dutch Agricultural Census), and theoretical proofs commonly published in major journals (e.g., EM algorithm for a Dirichlet distribution). It was helpful to see the thought process behind researchers and developers at Statistics Netherlands, given their leadership in the theory and practice of editing and imputation of survey data. One might notice that the handbook does not directly address the application of editing and imputation methodology in the context of non-probability sampling. In any event, the handbook would be a valuable resource for implementation of new editing and imputation programs in any agency. The thorough integration of theoretical, computational, and practical information covered the bases for management of edits or business rules.

References

Bankier, M. (1999). Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses. Working Paper No. 24. Rome: UN/ECE Work Session on Statistical Data Editing.

Fellegi, I. and Holt, D.T. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Johanson, J.M. (2012). Banff Automated Edit and Imputation on a Hog Survey. Proceedings of the Fourth International Conference of Establishment Surveys, June 11–14, 2012. Montréal, Canada [CD-ROM]: American Statistical Association.

Kosler, J.S. (2012). Survey Process Control with Significance Editing: Foundations, Perspectives, and Plans for Development. Proceedings of the Fourth International Conference of Establishment Surveys, June 11–14, 2012. Montréal, Canada [CD–ROM]: American Statistical Association.

Little, R.J. (2012). Calibrated Bayes, an Inferential Paradigm for Official Statistics. Journal of Official Statistics, 28, 309–334.

Robbins, M.W., Ghosh, S.K., and Habiger, J.D. (2012). Imputation in High Dimensional Economic Data as Applied to the Agricultural Resource Management Survey. Journal of the American Statistical Association (recently accepted for publication).

Joseph S Kosler, PhD
United States Department of Agriculture
National Agricultural Statistics Service (NASS)
Research and Development Division
3251 Old Lee Highway
Fairfax, VA 22030
E-mail: Joseph.Kosler@nass.usda.gov